THIS WEEK

EDITORIALS

WORLD VIEW Awareness of global warming could spread like wildfire **p.273**

LUCY'S LEGS Analysis of 3.2-million-year-old bones stands up **p.274**

ASTRONOMY Quasars find faint glow from star-free dark galaxies **p.275**

A different agenda

An attempt by Congress to save money by not funding political science seems to be motivated by ideological rather than financial reasons.

A fundamental question for democracy is what should be submitted to the democratic process. The laws of physics are presumably immune. But should public opinion help to decide which areas of science are studied or funded?

That is the implication of an amendment to the 2013 spending bill for the US National Science Foundation (NSF), which was approved by the House of Representatives in May. The amendment, proposed by Jeff Flake (Republican, Arizona), would prevent the NSF from funding political science, for which it awarded about US$11 million in grants this year. The Senate may well reject the amendment, but it is troubling that it has got so far, for two reasons.

First, to target a specific research programme marks an escalation from the familiar trick of finding research projects with apparently trivial titles and parading them as a waste of taxpayers' money. And second, scientists should ask themselves which vulnerable research programme could be next on the hit list — climate-change education, perhaps?

The social sciences are an easy target for this type of attack because they are less cluttered with technical terminology and so seem easier for the layperson to assess. As social scientist Duncan Watts at Microsoft Research in New York City has pointed out: "Everyone has experience being human, and so the vast majority of findings in social science coincide with something that we have either experienced or can imagine experiencing." This means that the Flakes of this world have little trouble proclaiming such findings obvious or insignificant.

Part of the blame must lie with the practice of labelling the social sciences as soft, which too readily translates as meaning woolly or soft-headed. Because they deal with systems that are highly complex, adaptive and not rigorously rule-bound, the social sciences are among the most difficult of disciplines, both methodologically and intellectually. They suffer because their findings do sometimes seem obvious. Yet, equally, the common-sense answer can prove to be false when subjected to scrutiny. There are countless examples of this, from economics to traffic planning. This is one reason that the social sciences probably unnerve some politicians, some of whom are used to making decisions based not on evidence but on intuition, wishful thinking and with an eye on the polls.

What of the critics' other arguments against public funding of political science? They say that the field is more susceptible to political bias; in particular, more social scientists have Democratic leanings than Republican. The latter is true, but it is equally so for US academics generally. We can argue about the reasons, but why single out political science? The charge of bias, meanwhile, is asserted rather than demonstrated.

So, what has political science ever done for us? We don't, after all, know why crime rates rise and fall. We cannot solve the financial crisis or stop civil wars, and we cannot agree on the state's role in systems of justice or taxation. As *Washington Post* columnist Charles Lane wrote in a recent article that called for the NSF not to fund any social science: "The 'larger' the social or political issue, the more difficult it is to illuminate definitively through the methods of 'hard science'."

In part, this just restates the fact that political science is difficult. To conclude that hard problems are better solved by not studying them is ludicrous. Should we slash the physics budget if the problems of dark-matter and dark-energy are not solved? Lane's statement falls for the very myth it wants to attack: that political science is ruled, like physics, by precise, unique, universal rules. In any case, we have little idea how successful political science has been — politicians rarely seem to pay much heed to evidence-based advice from the social sciences, unless of course that evidence suits them. And to constrain political scientists with utilitarian bean-counting undermines the free academic nature of the whole exercise.

"The idea that politicians should decide what is worthy of research is perilous."

The idea that politicians should decide what is worthy of research is perilous. The proper function of democracy is to establish impartial bodies of experts and leave it to them. But Flake's amendment does more than just disparage a culture of expertise. The research he selected for ridicule included studies of gender disparity in politics and models for international analysis of climate change — issues that are unpopular with right-wingers. In other words, his interference is not just about cost-cutting: it has a political agenda. The fact that he and his political allies seem to feel threatened by evidence-based studies of politics and society does not speak highly of their confidence in the objective case for their policies. Flake's amendment is no different in principle to the ideological infringements of academic freedom in Turkey or Iran. It has nothing to do with democracy. ■

Death of evidence

Changes to Canadian science raise questions that the government must answer.

The sight last week of 2,000 scientists marching on Ottawa's Parliament Hill highlighted a level of unease in the Canadian scientific community that is unprecedented in living memory.

The lab-coated crowd of PhD students, postdocs, senior scientists and their supporters staged a mock funeral for the 'death of evidence'. They said that the conservative government of prime minister Stephen Harper intends to suppress sources of scientific data that would refute what they see as pro-industry and anti-environment policies. Their list of alleged offences against science and scientific inquiry is lengthy and sobering.

It is important to note that the Harper government has increased science and technology spending every year since it took power in

2006, and has made a serious and successful attempt to attract top researchers to Canada. It has also set its sights on bolstering applied research, an area in which Canada has been relatively weak.

Nonetheless, the critics' specific complaints do give cause for deep concern — which is borne out by a close look at the specifics of the Harper budget that was passed into law late last month. In an effort to funnel more research money to commercialization and to erase the Canadian deficit by 2015, the government plans to cut the Research Tools and Instruments Grants Program (RTI), the main equipment-funding scheme for basic researchers, and to jettison the 24-year-old National Round Table on the Environment and the Economy (NRTEE), an independent source of expert advice to the government on sustainable economic growth. The government has also substantially weakened key laws that protect fish species and that require environmental assessments of development projects.

Of paramount concern for basic scientists is the elimination of the Can$25-million (US$24.6-million) RTI, administered by the Natural Sciences and Engineering Research Council of Canada (NSERC), which funds equipment purchases of Can$7,000–150,000. An accompanying Can$36-million Major Resources Support Program, which funds operations at dozens of experimental-research facilities, will also be axed. Canadian researchers have already warned the NSERC of 'drastic and irreversible' effects on the country's fundamental scientific research.

Even world-class facilities have not been spared. The government is closing the Polar Environment Atmospheric Research Lab (PEARL), located 1,100 kilometres from the North Pole and one of only three stations that keep a close watch on the polar atmosphere. The move comes just as data from the fast-changing Arctic climate are most needed. Another research station will be built to replace it, the government says, opening in 2017 — twice as far from the region it is supposed to monitor.

Equally disturbing is the proposed elimination next year of the internationally renowned Experimental Lakes Area (ELA) — a collection of 58 lakes and a field station in northwestern Ontario that has operated since 1968 as a natural laboratory. Work at the ELA has produced important evidence on the effects of acid rain and led to the discovery that phosphates from household detergents cause algal blooms. It has elucidated the impacts on fish of mercury and shown how wetland flooding for hydroelectricity leads to increased production of greenhouse gases.

It is hard to believe that finance is the true reason for these closures. PEARL costs the government about Can$1.5 million a year, and the ELA Can$2 million. The savings from eliminating the NRTEE would come to Can$5 million — all from a total science and technology budget of some Can$11 billion. Critics say that the government is targeting research into the natural environment because it does not like the results being produced.

"Scientific expertise and experience cannot be chopped and changed as the mood suits."

Instead of issuing a full-throated defence of its policies, and the thinking behind them, the government has resorted to a series of bland statements about its commitment to science and the commercialization of research. Only occasionally does the mask slip — one moment of seeming frankness came on the floor of the House of Commons in May, when foreign-affairs minister John Baird defended the NRTEE's demise by noting that its members "have tabled more than ten reports encouraging a carbon tax".

Governments come and go, but scientific expertise and experience cannot be chopped and changed as the mood suits and still be expected to function. Nor can applied research thrive when basic research is struggling. If the Harper government has valid strategic reasons to undermine vital sectors of Canadian science, then it should say so — its people are ready to listen. If not, it should realize, and fast, that there is a difference between environmentalism and environmental science — and that the latter is an essential component of a national science programme, regardless of politics. ■

London calling

The battle for gold is about to begin — and science is taking its place behind the podium.

As *Nature* went to press, excitement was mounting in the United Kingdom that Bradley Wiggins could become the first British cyclist to win the Tour de France this weekend. Win or lose, Wiggins will be back in the saddle a week or so later for the London Olympics — and he is already making headlines as he rebuffs Internet gossip that riders rely on performance-boosting drugs. "I cannot be doing with people [the critics] like that," was one of his more printable responses. "It justifies their own bone-idleness because they can't ever imagine applying themselves to do anything in their lives."

The use of drugs in sport and our inability to detect every case of misuse has an unfortunate side effect: the unfair suspicion that falls on those who win clean.

So, why bother? If we cannot ensure that everyone who competes is drug free, is one solution to remove the need for them to be so? That's one of a number of provocative ideas highlighted by a special series of Olympics-themed articles in this week's issue of *Nature*. (The opening ceremony next week, after all, will take place just a 5,000-metre race or so from our London headquarters.)

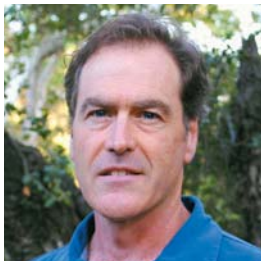
How much faster and stronger would an army of Olympians be if they were all allowed to get higher? And would medically supervised doping be safer? Some experts quoted in our News Feature on page 287 think so. One even goes so far as to call for a cross-sport 'pro-doping' agency to invest in safer forms of enhancement. And why stop at chemical help? The future could see runners with bionic limbs and swimmers with feet made webbed by skin grafts — developments that could demand separate events, so great would the advantages be.

It may sound far-fetched, but according to a Comment piece on page 297, the Olympic playing field is already tilted towards those with "unearned advantages" over the rest: their genes. Enough common genetic ground has been found to link successful athletes, the article says, to ask whether the Olympics is merely a showcase for "hardworking 'mutants'". If so, then would it be more sporting to hamper the lucky few — to make Usain Bolt run in heavy boots, say — or to cream the lot of them off into a separate competition entirely and leave the rest of us to have our mediocre fun?

For now, science remains a tool to catch those who break the existing rules, and to help those who want to play properly to compete. Profiles of some of the researchers who will work on these and other issues behind the scenes at London 2012 start on page 290. They include a psychologist who is working to assess intellectual disability in budding competitors in the Paralympics, and an epidemiologist waiting to map the inevitable spread of infectious disease among the several million expected visitors to London. Then there is the — unnamed — scientist who volunteered to be pulled through a swimming pool on a winch, subjected to a full body wax and then pulled through again, all to confirm what swimmers have long suspected, that body hair is a drag.

Finally, a Comment piece on page 295 examines the idea that humans evolved to run, and that a lifestyle without running could contribute to the modern boom in diseases such as obesity, diabetes and psychiatric disorders. Exercise doesn't just help muscles, it activates our brains. Armed with sticks and stones, our ancestors would have to chase down prey for hours, until the animals collapsed. The best weapon, the article says, was endurance. Bradley Wiggins, and the plucky researcher in the swimming pool, would surely agree. ■

➲ NATURE.COM
To comment online, click on Editorials at:
go.nature.com/xhunqv



S. COLE MORITZ

Wildfires ignite debate on global warming

As temperatures soar, forests blaze and houses burn, the media and public may be forced to face up to the reality of a changing climate, says **Max A. Moritz**.

I published an academic paper on climate change and global fire predictions last month, and I have been in my own media storm ever since. The huge wildfires that have broken out in the western United States had prompted dozens of enquiries from the press, nearly all asking the same question: "Are these fires due to climate change?".

For me, that marks a significant shift from previous years. During the conflagrations in southern California in 2003 and 2007, and the Black Saturday fires in the state of Victoria, Australia, in 2009, the question most reporters asked was: "Who is to blame here?".

This fresh curiosity about the link between fire and climate change is an important opportunity, of sorts. The media and the public seem to be searching for the evidence they need to take climate change more seriously. It is sad that it seems to take disasters to shift perspective, but perhaps they will also lead to a more science-based discussion of policy and planning. The term 'tipping point' gets thrown about too much, but I wonder if the United States is near one in terms of public perception about climate change.

The start of this year's fire season has been unusually fierce. Much of the western United States is extremely dry, and there are many reports of temperatures and forest fires that have broken records. The number of buildings destroyed — nearly 1,000 in a recent count — is staggering. Even if the fire season does not continue at the same a terrifying pace, these events could help to make climate change more real for many people. Is there a link with global warming? We have good reason to think so, and not taking the link seriously could have disastrous repercussions.

Climate change is not the only explanation. As usual, the conservative end of the political spectrum (including climate-change deniers) tends to blame environmental groups for opposing projects to thin forests, arguing that harvesting timber could have averted the devastating fires or mitigated their effects. Another argument focuses on the fact that we increasingly build homes in fire-prone ecosystems, including those that experience high-intensity fires as a natural event.

The latest fires in the interior west leave several open questions, and sometimes 'all of the above' is the best scientific explanation. Fire hazard can increase sharply after suppression of natural fires in dry forests of ponderosa pine, so the lack of active forest management (including prescribed fires) is indeed a potential culprit there. The picture is less clear for other forest types, and only further examination of fire-severity patterns will determine the role that forest management could have.

However, even if objections from environmentalists have contributed to more severe fires in some places, it does not follow that they contributed to the destruction of homes. Typically, structures ignite in exceptionally windy conditions, and this greatly offsets the effectiveness of forest thinning. Embers can be carried on the wind for kilometres until they find their way into a vulnerable spot, such as an unscreened vent or dry leaves under exterior decking. Poor planning decisions regarding building development and land use are at the heart of the structure-loss problem.

Most scientists avoid drawing conclusions about the contribution that climate change has made to forest fires on the basis of individual years or events. That said, the fires of this year and last seem to fit a documented pattern. Research shows a trend towards warmer spring and summer temperatures in many forests of the western United States, which leads to earlier melting of snow and a longer, more severe fire season.

The latest fires in the western United States are also consistent with models of fire activity expected from global-climate-change projections over the next few decades, including models that my lab helped to develop. The links to anthropogenic climate change are thus based on established relationships, operating at different scales of space and time, between climate and fire activity in various environments.

After reporters ask about wildfires and global warming, the next question is: "If these fires are related to climate change, what can we do about it?". Some people may cry "reduce greenhouse-gas emissions", but that is not what this question is about. Instead, these enquiries reveal a growing anxiety over how humanity can adapt to the fire-related impacts of climate change, rather than how to mitigate climate change itself.

To co-exist with fire will require extending our approach to living with environmental risks. Mapping other natural hazards, such as flood and earthquake zones, has taught us to avoid building on the most dangerous parts of the landscape or to engineer solutions into the built environment when we do. Encouraging the 'right kind of fire' — with frequencies, sizes and intensities appropriate to the ecosystem in question — will be necessary, where possible, so that 'record-breaking' fires are less likely to occur during 'record-breaking' heat or drought.

For some, climate change will become a fact only when its effects hit close to home. For this reason, perhaps we should expect an awareness of the need to adapt to climate change to precede a wider commitment to mitigating climate change itself. If that is the case, reporters are, finally, asking the right questions. ■

FOR SOME, CLIMATE CHANGE WILL BECOME A **FACT** ONLY WHEN ITS EFFECTS HIT CLOSE TO **HOME.**

NATURE.COM
Discuss this article online at:
go.nature.com/obuvxy

Max A. Moritz *is a faculty member in the Environmental Science, Policy, and Management Department at the University of California, Berkeley.*
e-mail: mmoritz@berkeley.edu

RESEARCH HIGHLIGHTS

Selections from the scientific literature

NEUROSCIENCE

Hormone linked to depression

A hormone released by fat cells that is associated with a reduced risk of type 2 diabetes could also protect against depression.

Blood levels of the hormone adiponectin are positively correlated with insulin sensitivity. Xin-Yun Lu at the University of Texas at San Antonio and her colleagues manipulated levels of adiponectin in the blood and brains of mice, and assessed the animals' likelihood of exhibiting depression-like behaviours in response to stressors. The researchers found that mice with low blood levels of the hormone were more likely to show signs of depression than were control animals. Injecting adiponectin-neutralizing antibodies into the mouse brain also increased the likelihood of depression-like symptoms. By contrast, injecting adiponectin into the brain had an antidepressant-like effect in both normal-weight and obese, diabetic mice.

The findings could explain why depression is twice as prevalent in people with type 2 diabetes as in the general population.

Proc. Natl Acad. Sci. USA **http://dx.doi.org/10.1073/pnas.1202835109 (2012)**

A. DESCAT/MAP/GARDEN WORLD

ECOLOGY

For more than just kissing

The ecological impact of mistletoe plants — proposed to be keystone species crucial to ecosystem health — has been quantified. This is the first such attempt for any keystone species.

Mistletoes (Loranthaceae; pictured) provide fruit and nectar, as well as nesting and roosting sites for animals such as birds and insects. David Watson and Matthew Herring at Charles Sturt University in Albury, Australia, removed mistletoes from 17 woodland sites in the state of New South Wales in 2004. Three years after the plants' removal, species richness had declined by an average of 20.9% overall, and by 26.5% in the case of woodland-dependent bird species. By contrast, in 11 control sites where mistletoes remained, average species richness had increased.

Mistletoes promote biodiversity mainly by enriching the soil with nutrients through leaf fall and decomposition, the authors suggest.

Proc. R. Soc. B. **http://dx.doi.org/10.1098/rspb.2012.0856 (2012)**

PALAEOANTHROPOLOGY

Lucy's relatives walked upright

An analysis of bones from the same species as 'Lucy' — a hominin who lived 3.2 million years ago — suggests that this species was more human-like than previously thought.

C. WARD

Carol Ward at the University of Missouri in Columbia and her team analysed dozens of *Australopithecus afarensis* bones (**example pictured**), unearthed between 1990 and 2007 in Hadar, Ethiopia. The bones have been dated to between 3 million and 3.4 million years ago, and are thought to be from individuals intermediate in size between the smallest and the largest *A. afarensis* specimens yet found.

Foot bones indicate that *A. afarensis* had an arched foot, whereas vertebrae suggest a more human-like upper backbone than previously suspected. Both are consistent with upright walking.

J. Hum. Evol. **http://dx.doi.org/10.1016/j.jhevol.2011.11.012 (2012)**

MICROBIOLOGY

Watching biofilms form

A sophisticated imaging technique has enabled researchers to watch bacteria assemble into tight-knit, organized communities called biofilms and to discern the structure and key protein components that hold these communities together.

Biofilms help bacteria to survive stressors such as antibiotics, but studying intact, living biofilms has proved difficult. Veysel Berk at the University of California, Berkeley, and his team developed a method to fluorescently label proteins in cells and continuously image them using conventional and super-resolution microscopy. The researchers watched dividing *Vibrio cholerae* cells, which cause cholera, and found that biofilms form when daughter cells remain attached to their parent cells, generating cell clusters. These clusters group together and are enclosed by a protein envelope to ultimately form the biofilm.

The team also pinpointed the roles of four proteins that form the protein envelope and that allow cells to stick to each other and to surfaces.
Science 337, **236–239 (2012)**

ASTRONOMY

Dark galaxies revealed

'Dark' galaxies contain no stars, making them impossible to observe using optical telescopes. Now, Sebastiano Cantalupo and his colleagues at the University of California, Santa Cruz, have managed to detect a faint fluorescent glow from a few such galaxies.

The researchers used the European Southern Observatory's Very Large Telescope to look at bright galaxies known as quasars, which can illuminate nearby dark galaxies. The team was able to see the glowing outline of a dozen candidate dark galaxies, thanks to radiation from a quasar exciting the hydrogen gas in these galaxies.

The team estimates the dark galaxies contain a gas mass around one billion times that of the Sun. These galaxies could serve as reservoirs of hydrogen fuel for star formation in larger galaxies, the authors suggest.
Mon. Not. R. Astron. Soc. **http://dx.doi.org/10.1111/j.1365-2966.2012.21529.x (2012)**

COGNITIVE NEUROSCIENCE

A smart hub in the brain

Human intelligence could result from high levels of activity in a region of the brain called the lateral prefrontal cortex, as well as strong connectivity between this region and the rest of the brain.

Human intelligence is marked by the ability to direct thoughts and behaviour in pursuit of a goal. This 'cognitive control' has previously been linked to activity in the lateral prefrontal cortex. Using functional magnetic resonance imaging, Michael Cole at Washington University in Saint Louis, Missouri, and his team examined the activity of this region and its connectivity across the brain in 94 young adults. Individuals with high levels of connectivity were more likely to perform well on tests of both cognitive control and intelligence.

The lateral prefrontal cortex could act as a global hub for human intelligence that exerts its effects across the brain.
J. Neurosci. 32, **8988–8999 (2012)**

BIOPHYSICS

Trout nose yields magnetic cells

Certain animals, including some birds and fish, are guided by magnetic fields, and researchers have isolated magnetic cells that could be at the root of this internal compass.

Michael Winklhofer at Ludwig-Maximilians-University in Munich, Germany, and his colleagues took epithelial cells from the rainbow trout nose and exposed them under a microscope to a moderately strong, rotating magnetic field. A few of the cells spun at the same frequency as the magnetic field, indicating that they were sensitive to the field, whereas the other cells did not alter their behaviour. High-resolution imaging showed that the responsive cells had micrometre-scale structures composed of iron-rich crystals attached to their cell membranes. These caused the cells to align with the magnetic field.
Proc. Natl Acad. Sci. USA **http://dx.doi.org/10.1073/pnas.1205653109 (2012)**

BIODIVERSITY

Extinctions still to come

Up to 90% of extinctions due to Amazon rainforest loss have yet to occur, a modelling study suggests.

Robert Ewers and his team at Imperial College London in Ascot used deforestation data from 1978 to 2008 — plus data on forest-dependent vertebrates — to create a model that relates species extinction to the timing and amount of habitat loss. The model takes into account the fact that species do not become extinct immediately after losing their habitats and that deforestation happens intermittently, rather than all at once as previous models have assumed.
The authors calculated the number of species headed for extinction owing to previous deforestation, or the 'extinction debt': an average of two mammals, four or five birds, and one amphibian for every 2,500 square kilometres.

The team then projected Amazon extinctions up to 2050 on the basis of four scenarios of different levels of forest regulation. They found that in the most likely scenario, 60–70% of expected extinctions would be yet to come as a result of past and future habitat loss.
Science 337, **228–232 (2012)**
For a longer story on this research, see http://go.nature.com/rmo2d3

A. LEES

COMMUNITY CHOICE

The most viewed papers in science

CANCER BIOLOGY

p53 can be cancer's friend, not foe

HIGHLY READ on www.cell.com 10 June–10 July

The protein p53 is well known for its role as a tumour suppressor; however, p53 can also help breast tumour cells to dodge the effects of chemotherapy.

Normal p53 activates programmed cell death, or apoptosis, and the p53 gene is often mutated in cancer. Researchers expected that tumours with mutant p53 would be more difficult to treat than those with normal p53, but previous studies on breast cancer have not found this. Guillermina Lozano at the MD Anderson Cancer Center in Houston, Texas, and her team therefore tested the effects of a chemotherapy drug, doxorubicin, on mice bearing breast tumours with either normal or mutated p53.

Tumours in mice with normal p53 shrank little in response to treatment and relapsed more quickly than those with mutated p53. After treatment, many tumour cells with normal p53 entered a state of senescence in which they stopped dividing — however, the cells secreted signalling proteins that could trigger the proliferation of neighbouring cells, leading to relapse.
Cancer Cell 21, **793–806 (2012)**

NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

POLICY

FDA spy saga

The US Food and Drug Administration (FDA) conducted a vast surveillance operation of five employees it believed were leaking confidential information, it emerged this week. The agency used computer software to gather more than 80,000 pages of records tracking the scientists' communications with politicians, lawyers, journalists and US president Barack Obama. *The New York Times* exposed the scope of the operation on 14 July after the cache of records was inadvertently posted on a public website by an FDA contractor. The scientists, four of whom have been fired, are suing the FDA, alleging that their e-mails were monitored. They had earlier raised concerns that the agency was approving unsafe medical imaging devices (see *Nature* **482,** 136; 2012).

Research integrity

Major science funders in the United Kingdom are introducing a set of principles on research integrity as a condition for receiving grants. *The Concordat to Support Research Integrity*, which includes commitments to transparency and rigour in research, was launched on 11 July. It stipulates that those who employ researchers must have clear and confidential mechanisms for reporting allegations of misconduct, and must provide annual summaries of their activities in this area, including statements on any formal investigations. See go.nature.com/f7smde for more.

Open access

The United Kingdom announced on 16 July that its publicly funded research would be made free to read. From April 2013, research findings paid for by the country's seven research councils (government-funded grant agencies) must be free to access within six months of publication, Research Councils UK said. The next day, the European Commission announced similar ambitions in proposals for its 2014–20 research-funding programme, Horizon 2020. See page 285 for more.

J. LEVAC/*OTTAWA CITIZEN*

Scientists march on Canada's parliament

Some 2,000 researchers marched through the streets of Ottawa to Canada's Parliament Hill on 10 July, to protest against severe cuts to government research labs and — the scientists say — a lack of evidence-based decision-making in the Canadian government. Cuts to the federal budget this year have meant the closure of various scientific programmes, including the Experimental Lakes Area, a 44-year-old research station encompassing a system of 58 freshwater lakes in northwestern Ontario. See go.nature.com/dbz3bm for more.

Telescope club full

The Large Synoptic Survey Telescope (LSST), which aims to map the southern sky in unprecedented detail, now has a full complement of international partners. The success increases the chances that the US National Science Foundation, a key sponsor, will approve the project, allowing it to begin operations in Chile in 2022. See page 284 for more.

HIV prevention

The US Food and Drug Administration on 16 July approved the drug Truvada, a combination of the antiretroviral drugs emtricitabine and tenofovir, as a way to reduce the risk of sexually acquired HIV infection. It is the first medication approved to prevent the disease. See page 283 for more.

Biosafety bumped

Opponents of plans to build a National Bio- and Agro-Defense Facility (NBAF) in Kansas are cheering a review released on 13 July by the US National Research Council. The NBAF would be the only US lab able to study diseases in cows and horses at the highest biosafety level, but the idea of building it in America's 'cattle country' had provoked protests. The report affirms the need for such facilities, but suggests scaling back the plans for the NBAF and moving some its functions to existing labs. See go.nature.com/rruzdy for more.

Controversial prize

After years of argument, the United Nations Educational, Scientific and Cultural Organization (UNESCO) was set to award a prize for life-sciences research sponsored by Teodoro Obiang Nguema Mbasogo, president of

Equatorial Guinea, on 17 July. The prize was proposed in 2008 but had been in limbo as a result of opposition from Western diplomats and UNESCO's director-general, Irina Bokova, who pointed to corruption and human-rights abuses in Equatorial Guinea. But in March, UNESCO's executive board narrowly voted to remove Obiang's name from the award's title and push ahead. See go.nature.com/dkekrr for more.

RESEARCH

Pluto's fifth moon

The Hubble Space Telescope has discovered another moon orbiting the dwarf planet Pluto, NASA announced on 11 July. P5, as the moon is informally known, is just 10–25 kilometres across, which is smaller than the satellites P4 (found last July), and Nix and Hydra (spotted in 2005). These are all tiny compared with Pluto itself (around 2,300 kilometres in diameter) or its largest moon Charon (1,200 kilometres). See go.nature.com/6kpdse for more.

Atmospheric lab

Plans to cut the number of scientists and reduce measurements at a world-class atmospheric-research centre in New Zealand prompted a barrage of international concern last

NIWA

week. The 51-year-old Lauder Atmospheric Research Station (**pictured**) on the South Island specializes in measuring levels of chlorofluorocarbons, ultraviolet light and greenhouse gases. But New Zealand's National Institute of Water and Atmospheric Research, the government-owned company that administers the lab, has told Lauder staff that it plans to axe all three of the site's atmospheric-scientist positions. See go.nature.com/jzw1dm for more.

Pathogen genomes

An open-access database of the genomes of 100,000 food-borne pathogens will be built over five years in an effort to speed up identification of disease outbreaks and development of diagnostic tests. On 12 July, the US Food and Drug Administration (FDA) announced the 100K Genome Project, which will sequence proven bacterial pathogens provided by the FDA and the Centers for Disease Control and Prevention in Atlanta, Georgia. The project will be headquartered at the University of California, Davis, with support from Agilent Technologies, based in Santa Clara, California.

BUSINESS

Biotech buyout

British drug giant GlaxoSmithKline (GSK) announced on 16 July that it had acquired the biotech firm Human Genome Sciences, based in Rockville, Maryland in a deal worth US$3.6 billion. GSK has been pursuing the firm for some months; a $2.6-billion offer was rejected in April. Together, the two companies developed Benlysta (belimumab) for systemic lupus and have candidates for heart disease (darapladib) and diabetes (albiglutide) in clinical trials. GSK will now own all three drugs.

PEOPLE

Data detective

Uri Simonsohn, the researcher who flagged up questionable data in studies by social psychologist Dirk Smeesters, last week revealed to *Nature* that he believes data from a second social psychologist, Lawrence Sanna, is suspiciously perfect. Sanna's former employer, the University of Michigan in Ann Arbor, says that he resigned his professorship there at the end of May. It is not clear why Sanna resigned — but his departure followed questions from Simonsohn and a review by Sanna's previous institution, the University of North Carolina at Chapel Hill. Sanna has asked that three of his papers be retracted from the *Journal of Experimental Social Psychology*. See go.nature.com/fgfzvi for more.

DARPA director

Arati Prabhakar will be the next chief of the Pentagon's research arm, the US Defense Advanced Research Projects Agency (DARPA) in Arlington, Virginia. Prabhakar, who trained as a physicist and headed the National Institute of Standards and Technology before working in venture-capital investment, starts on 30 July. She replaces Regina Dugan, the agency's first female director, who in March left after two years to work for Google.

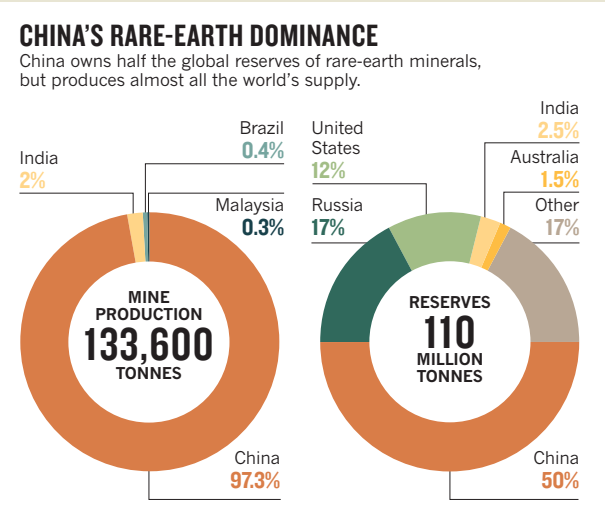
COMING UP

22–27 JULY
Latest results on efforts to cure HIV are presented at the International AIDS Conference in Washington DC.
www.aids2012.org

27 JULY
Chemist Patrick Harran and the regents of the University of California are called before court to answer criminal charges over the death of 23-year-old Sheharbano Sangji in a laboratory fire 3.5 years ago.
go.nature.com/cvxyii

TREND WATCH

Japan and Vietnam last month launched the joint Rare Earth Research and Technology Transfer Centre in Hanoi, part of efforts to break China's monopoly on rare-earth elements (see graphs). The 17 metals are used in high-tech applications such as catalysts, but prices have rocketed as China has imposed export limits. The centre aims to develop technologies to separate and concentrate the elements, which would enable processing of ores from more mines worldwide. See go.nature.com/cebfoj for more.



SOURCE: USGS

NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS

R. BUENDIA/AFP/GETTY

Lonesome George, who died last month, was the last of the Pinta giant tortoises.

CONSERVATION

The legacy of Lonesome George

Tortoise's death spurs Galapagos conservation efforts.

BY HENRY NICHOLLS IN PUERTO AYORA, GALAPAGOS

Even in death, Lonesome George's star power burns brightly. After the iconic giant tortoise died last month, Ecuadorian President Rafael Correa mourned the reptile's loss in an address to the nation, expressing hope that "one day, science and technology will be able to reproduce him, to clone him".

George was the last of the Pinta tortoises (*Chelonoidis abingdoni*), and it is too soon to know whether biotechnology can reverse the extinction. But scientists have made the first step: in the days after George's death, they raced to keep his cells alive, collecting tissue and ferrying liquid nitrogen to his remote home in Ecuador's Galapagos Islands to preserve samples that might one day yield a viable cell culture.

Even if the bid fails, George's death is already offering hope for other giant tortoises. Last week, *Nature* joined experts in Puerto Ayora on the island of Santa Cruz for an international workshop dedicated to the memory of Lonesome George. The meeting aimed to galvanize efforts to prevent the loss of other Galapagos tortoise species and their habitats. "One species is very important, but most important are the ecosystems," says Washington Tapia, director of conservation and sustainable development for the Galapagos National Park.

George was discovered on the island of Pinta in 1971, and was transferred to the Charles Darwin Research Station in Puerto Ayora the following year. Conservationists launched a long and frustrating campaign to persuade the reptile to mate with females from other Galapagos islands[1]. Hopes soared in 2008 when handlers discovered eggs in the enclosure he shared with two females, but the eggs were found to be unfertilized[2].

On 24 June this year, Fausto Llerena, a ranger at the Galapagos National Park and George's long-term keeper, found him slumped in his corral. Within hours, Llerena was helping to carry the tortoise's corpse, trussed onto a wooden frame, into a chilled storage chamber.

In 2008, Ecuador was the first country in the world to amend its constitution explicitly to grant basic rights to nature and its inhabitants. As a consequence, George was accorded a full necropsy, performed by Marilyn Cruz, a veterinary surgeon who is the coordinator of Agrocalidad in Galapagos, the government agency that oversees agriculture and biosecurity on the islands. "The last thing we need to do is to investigate his tissues," says Cruz, currently the legal guardian of George's remains.

Cruz found nothing obviously wrong with the tortoise; she concluded that he probably died of natural causes. But, she says, George's "liver and kidneys appear to have some abnormalities", which the laboratories need to investigate in depth. Cruz also took a sample of George's skin cells for culturing; they could eventually be used to generate stem cells and sex cells, or in reproductive cloning.

But the cells had to be frozen within days of George's death. Establishing a viable cell culture "is highly correlated with the freshness of the sample", explains Oliver Ryder, a geneticist at San Diego Zoo in California and champion of the Frozen Zoo, a facility containing cryopreserved ▶

NATURE.COM
To read a News Feature on turtle conservation, see: go.nature.com/merukq

cell cultures from more than 9,000 individuals representing nearly 1,000 endangered species. Ryder arranged for colleagues in San Diego to deliver tissue-culture medium and the cryoprotectant dimethyl sulphoxide to Ecuador on the first available flight.

Cruz hurried to secure a local supply of liquid nitrogen. There were two liquid nitrogen tanks on Santa Cruz, used for freezing semen for the artificial insemination of cattle in the highlands. By commandeering one tank, Cruz got hold of enough liquid nitrogen to keep George's cells safely frozen until top-up stock could be shipped from the mainland. Several liquid-nitrogen containers are now being relayed back and forth between the continent and the islands to keep the cells frozen until it is decided where they will be stored in the long-term. It remains to be seen whether the cells, if thawed, will yield a viable culture.

With George's demise, ten species of Galapagos tortoise remain. The reptiles' populations have suffered as a result of hunting, habitat destruction and the introduction of destructive species over the past few hundred years; some are now on the increase as a result of conservation efforts, but with a tortoise typically taking 20–30 years to reach sexual maturity, recovery has been very slow.

Last week's workshop took several years to organize, and one subject on the agenda was what to do with George in the event of his death, and whether to collect his cells while he was still alive. With that no longer possible, the workshop focused on its main goal of thrashing out a ten-year plan to preserve the surviving animals (see 'Endangered Galapagos giants').

"What we're trying to do is bring everyone together" to synthesize the perspectives of ecologists, geneticists and conservation managers, says workshop organizer Linda Cayot, science adviser to the Galapagos Conservancy in Fairfax, Virginia. The outcome, she says, will be a single set of recommendations that can be delivered to the Galapagos National Park.

The meeting also began work to review the status of the Galapagos tortoises on the International Union for Conservation of Nature (IUCN) Red List of threatened species. The tortoises' Red List entries date from 1996 and are in urgent need of revision, says Peter Paul van Dijk, co-chairman of the IUCN/Species Survival Commission Tortoise and Freshwater Turtle Specialist Group, who attended the meeting. Some Galapagos tortoises could have their threat categories downgraded. But the Pinta tortoise — still listed as 'Extinct in the Wild' — will be recategorized as 'Extinct'.

To some visitors, the giant tortoises on one island of the Galapagos might look much the same as those on another. But as Charles Darwin came to appreciate after his brief sojourn in the Galapagos in 1835, each island or main volcano seems to have its own distinct type of tortoise, and all are diverging into separate species. Genetic differences suggest that Lonesome George's own ancestors somehow travelled to Pinta from the island of Española about 300,000 years ago, and had been diverging from their relatives ever since.

From a management perspective, "each island is totally different", says Cayot, who was one of the first researchers to carry out an in-depth study[3] of the behavioural ecology of giant tortoises, in the early 1980s. "Pinzón has rats. Santiago had pigs and goats. Pinta had goats, but only for 20 years. Española had goats for probably 100 years," she says. "That's what makes Galapagos so much fun."

One of the most fascinating populations lives around Wolf volcano at the northern tip of the island of Isabela. Over the past decade, Adalgisa Caccone, a geneticist at Yale University in New Haven, Connecticut, and her colleagues have been unpicking the ancestry of this mixed-up population. In a study[4] based on blood samples from a few dozen individuals, she found evidence that tortoises from Española and San Cristóbal had crossed more than 250 kilometres of sea to reach Wolf, probably carried by pirates and whalers. Using DNA from museum specimens, the researchers went on to show[5,6] that Wolf is also harbouring descendants of the long-lost Floreana lineage and the recently lost Pinta one.

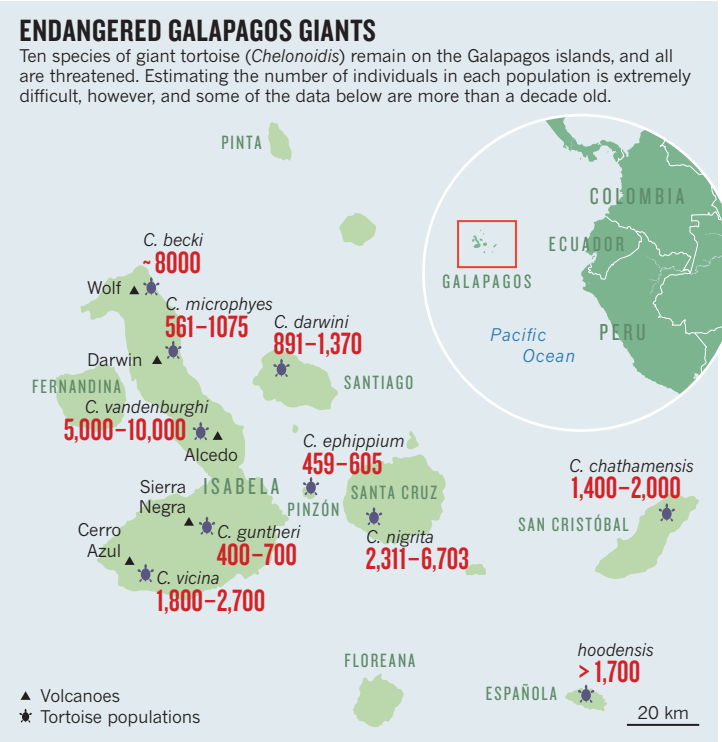
The researchers hope to mount a return expedition to Wolf volcano next year, in an effort to locate the Floreana- and Pinta-like tortoises. In theory, these animals could be taken off Wolf volcano for captive breeding.

Floreana has been heavily affected by habitat destruction and introduced species, and has been without tortoises for more than 150 years. But the Floreana-like tortoises on Wolf could help with a long-term project to restore the island's ecology. The situation on Pinta is more urgent, and waiting for a captive-breeding programme to bear fruit may not be an option. Much of the island's original vegetation is intact, but without tortoises — once the island's dominant herbivore — there is a danger that some plant species could be choked out and lost.

If a rapid solution cannot be found using tortoises of Pinta pedigree, it looks increasingly likely that conservationists will introduce a species from another island. "Given that tortoises from Española founded the original population that landed on Pinta and evolved into the Pinta tortoises, I don't see a problem with us repopulating that island with Española tortoises," says Cayot.

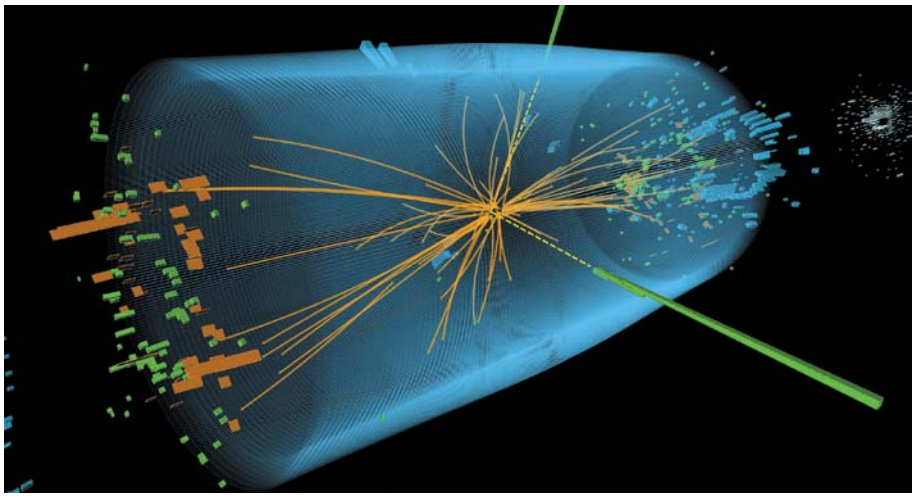
The Española tortoise was once on the brink of extinction, but now there are more than 1,700 of the reptiles, and conservationists can afford to consider transferring some of them to Pinta. This kind of deliberate introduction is unprecedented in the Galapagos, however, so researchers are cautious. As a precursor experiment, almost 40 sterilized hybrid tortoises have been introduced to Pinta, and are being tracked by satellite to see what impact they have on the ecosystem.

For Cayot, introducing a breeding population of tortoises to Pinta is a much more rational proposal than a plan that relies on cloning Lonesome George. "In 100,000 years, through evolutionary processes, we'll have a Pinta tortoise in Galapagos," she says. "100,000 years is a time frame I can deal with." ■



SOURCE: C. MÁRQUEZ *ET AL. ECOLOGIA APLICADA* **3,** 98–111 (2004)/GALAPAGOS NATL PARK

1. Nicholls, H. *Nature* **429,** 498–500 (2004).
2. Nicholls, H. *Nature* http://dx.doi.org/10.1038/news.2008.1221 (2008).
3. Cayot, L. J. *Ecology of giant tortoises* (Geochelone elephantopus) *in the Galapagos Islands.* PhD Thesis, Syracuse Univ. (1987).
4. Caccone, A. *et al. Evolution* **56,** 2052–2066 (2002).
5. Poulakakis, N. *et al. Proc. Natl Acad. Sci. USA* **105,** 15464–15469 (2008).
6. Russello, M. A. *et al. Curr. Biol.* **17,** R317–R318 (2007).



MCCAULEY; L. TAYLOR/CERN

The Higgs boson seems to decay into two photons more often than particle physicists expected.

PARTICLE PHYSICS

Theorists feast on Higgs data

But usurpers of 'standard model' have little to chew on.

BY GEOFF BRUMFIEL

The popping of champagne corks may have subsided since scientists presented convincing evidence for the existence of the long-sought Higgs boson on 4 July[1], but the work has just begun for theoretical particle physicists, who are revelling in the biggest glut of data they've had since the 1990s. Many are working evenings and weekends to interpret the results, and they have already generated a publication boom, with dozens of papers about the discovery appearing on the preprint server arXiv.org during the past two weeks.

Some are using the fresh data from the Large Hadron Collider (LHC) at CERN, Europe's particle-physics laboratory near Geneva in Switzerland, to eliminate theoretical models. Others are probing for hints of new particles. Most are still hoping that their investigations will produce a grand theory to replace the almost infallible standard model of particle physics, a framework that predicts the behaviour and properties of all fundamental particles and every force except gravity. Despite its success, the standard model contains some mathematical chinks hinting that there must be a deeper truth about how the Universe works — and theorists the world over dream of finding it first.

However, for those who have spent their careers pursuing a more powerful extension of the standard model called supersymmetry (SUSY), the data offer scant succour. The theory predicts a suite of particles that are 'super-partners' to all the known particles, along with several types of Higgs boson. Many theorists regard SUSY as the most promising route to a broader theory of particles and forces, and a possible solution to puzzles such as the nature of cosmic dark matter.

But the LHC has yielded few signs of SUSY. Aside from a handful of tantalizing observations, the Higgs boson seems to match the standard model's predictions perfectly. Under the weight of the LHC's hard evidence, SUSY and other beloved theories are feeling the strain. "There's going to be a huge massacre of theoretical ideas in the next couple of years," predicts Joe Lykken, a theoretical physicist at Fermilab in Batavia, Illinois.

The Higgs is often described as the last missing piece of the standard model. It is a manifestation of the Higgs field, which is thought to confer mass on the other known particles. The results presented earlier this month measured the mass of the Higgs boson at about 125 gigaelectronvolts (GeV), roughly 133 times the mass of a proton. Filling in that part of the particle-physics jigsaw is a triumph, but it also exposes a fundamental vulnerability of the standard model. Its equations predict that as the Higgs interacts with other particles, the boson's mass should fluctuate wildly. That researchers were able to pin down its mass to 125 GeV suggests that some stabilizing mechanism is at play; and theorists have previously invoked extra dimensions, new forces and exotic particles, including those predicted by SUSY, to fix the problem.

NATURE.COM
For all of *Nature*'s LHC coverage, see:
go.nature.com/uzvtfd

The latest data seem to rule out some of the more left-field possibilities. The very existence of the Higgs knocks out a Higgs-free version of the standard model known as Technicolor, for example. And hopes of finding extra dimensions that would mysteriously swallow up energy from collisions in the LHC are evaporating faster than the postulated microscopic black holes that also failed to make an appearance. "I was one of the people who pushed the idea of extra dimensions that we could see in our lifetime," says Lykken. "Now that we have data, I'm becoming much more conservative." Lykken has turned to the possibility that an exotic 'imposter' particle is creating a phantom Higgs signal inside the machine[2].

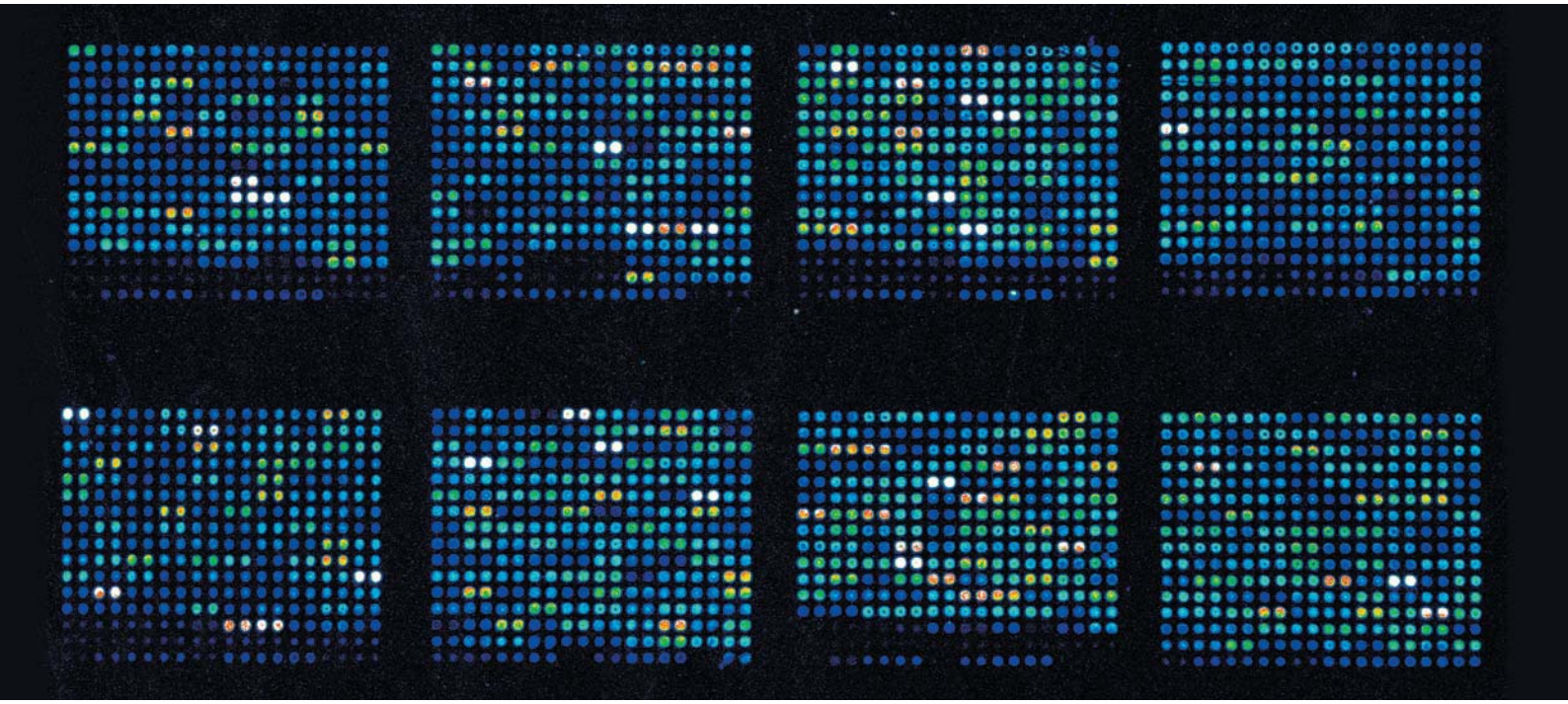
Many of the recent arXiv papers focus on the slivers of hope for SUSY that are buried in the new data. The Higgs particle found at CERN seems to be decaying into pairs of photons about twice often as expected, potentially a sign that it is interacting with a super-partner of the Top quark, according to Fermilab's Dan Hooper[3]. If the signal grows stronger as more data are collected, it could offer the first evidence of SUSY's extra particles.

But the mass of the Higgs particle is too large to be explained by the simplest, or 'minimal', supersymmetry models. So some theorists are moving on to greener pastures outside the LHC's confounding reach, developing 'near-minimal' versions of the theory that could explain both the heavier mass of the Higgs and the apparent absence of super-particles. It will take years' more data to test some of the most promising ideas, says Gordon Kane, a theorist at the University of Michigan, Ann Arbor, and a longtime SUSY champion[4].

It is too soon to write off SUSY, agrees Frank Wilczek, a physicist at the Massachusetts Institute of Technology in Cambridge who was awarded the Nobel Prize in Physics in 2004 for his work on the standard model. "The last man standing, as far as ambitious ideas beyond the standard model go, is supersymmetry."

But there's a growing sense that SUSY is being painted into a corner. "If you're the optimistic type, you could look to these small deviations and hope they can grow," says Tomer Volansky, a theorist at Tel Aviv University. He is setting aside dreams of replacing the standard model while he concentrates on working out how closely the Higgs' decay patterns match predictions[5]. "Right now," he says, "I want to listen to the data rather than thinking about theories." ■

1. Brumfiel, G. *Nature* **487,** 147–148 (2012).
2. Low, I., Lykken, J. & Shaughnessy, G. preprint at http://arxiv.org/abs/1207.1093 (2012).
3. Buckly, M. R. & Hooper, D. preprint at http://arxiv.org/abs/1207.1445 (2012).
4. Kane, G. *Nature* **480,** 415 (2011).
5. Carmi, D., Falkowski, A., Kuflik, E., Volansky, T. & Zupan, J. preprint at http://arxiv.org/abs/1207.1718 (2012).



DNA microarrays allow researchers to analyse the expression of a huge number of genes simultaneously.

GENOMICS

Gene data to hit milestone

With close to one million gene-expression data sets now in publicly accessible repositories, researchers can identify disease trends without ever having to enter a laboratory.

BY MONYA BAKER

Purvesh Khatri sits in front of an oversized computer screen, trawling for treasure in a sea of genetic data. Entering the search term 'breast cancer' into a public repository called the Gene Expression Omnibus (GEO), the postdoctoral researcher retrieves a list of 1,170 experiments, representing nearly 33,000 samples and a hoard of gene-expression data that could reveal previously unseen patterns.

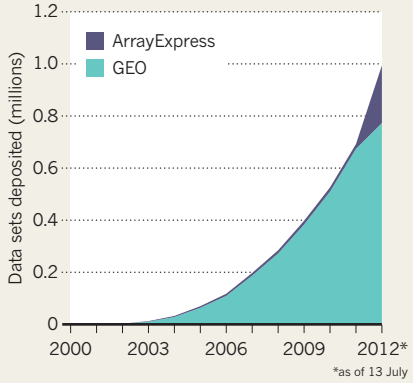
That is exactly the kind of search that led Khatri's boss, Atul Butte, a bioinformatician at the Stanford School of Medicine in California, to identify a new drug target for diabetes. After downloading data from 130 gene-expression studies in mice, rats and humans, Butte looked for genes that were expressed at higher levels in disease samples than in controls. One gene was strikingly consistent: *CD44*, which encodes a protein found on the surface of white blood cells, was differentially expressed in 60% of the studies (K. Kodama *et al. Proc. Natl Acad. Sci. USA* **109,** 7049–7054; 2012). The CD44 protein is not widely investigated as a drug target for diabetes, but Butte's team found that treating obese mice with an antibody against it caused their blood glucose levels to drop.

Butte and his team are now using publicly available data to answer a diverse range of questions — Khatri, for instance, hopes to discover secrets behind kidney-transplant rejection. "We don't do wet lab experiments for discovery," he says. Those are for validating hypotheses. The beauty of analysing data from multiple experiments is that biases and artefacts should cancel out between data sets, helping true relationships to stand out, Butte says. "There is safety in numbers."

And those numbers are rising rapidly. Since 2002, many scientific journals have required that data from gene-expression studies be deposited in public databases such as GEO, which is maintained by the National Center for Biotechnology Information in Bethesda, Maryland, and ArrayExpress, a large gene-expression repository at the European Bioinformatics Institute (EBI) in Hinxton, UK. Some time in the next few weeks, the number of deposited data sets will top one million (see 'Data dump').

DATA DUMP

The number of gene-expression data sets in publicly available databases has climbed to nearly one million over the past decade.



SOURCES: NIH, EBI

The result is an unprecedented resource that promises to drive down costs and speed up progress in understanding disease. Gene-sequence data are already shared extensively, but expression data are more complex and can reveal which genes are the most active in, say, liver versus brain cells, or in diseased versus healthy tissue. And because studies often look at many genes, researchers can repurpose the data sets, asking questions other than those posed by the original researchers.

It is easy to track how many data sets are being deposited — much harder is working out how they are being used. Heather Piwowar, who studies data reuse with the National Evolutionary Synthesis Center from the University of British Columbia in Vancouver, Canada, found that 20% of data sets deposited in GEO in 2005 and 17% of those in 2007 had been cited by the end of 2010. But those rates are certainly underestimates, she says. The PubMed Central repository, which her study relied on, holds only about one-third of the relevant papers, and her algorithms identify reuse only when researchers cite database accession numbers, which many don't do. More studies are reusing data every year, she says. "We have every reason to believe it is game-changing."

Having access to such data is "immensely

A. NANTEL/SHUTTERSTOCK

valuable," agrees Enrico Petretto, a genomicist at Imperial College London. "We would never be in a position to look across multiple tissues and species with the money we have." But he cautions that using other people's data can be tricky. If data sets give contradictory outcomes, it is unclear whether that is because the underlying data contradict each other or because something went wrong with the analysis. "That's why people sometimes don't trust this," he says.

CHANGE OF PRACTICE

Still, few researchers are using the data to their greatest potential, says Alvis Brazma, a bioinformatician at the EBI. "Being able to reuse functional genomics data is a really new thing," he says. Researchers rarely download more than half a dozen data sets, and most use the data only to compare with their own results. Studies that use only other scientists' data to come up with new findings are still unusual.

That makes Butte and Khatri trailblazers. Another pioneer is Gustavo Stolovitzky, a computational biologist at the IBM Thomas J. Watson Research Center in Yorktown Heights, New York, who has used publicly available data to train algorithms to recognize gene signatures for diseases such as lung cancer, chronic obstructive pulmonary disease (COPD) and psoriasis. Not only can the algorithms distinguish lung cancer from COPD, they can also tell squamous-cell carcinoma from adenocarcinoma. "There is enough info in existing databases to predict disease in samples that algorithms have never seen before," Stolovitzky says.

Other efforts promise to unleash even more power from the growing repositories. In 2009, for instance, curators of ArrayExpress used their database to create the Gene Expression Atlas, which allows researchers to look at how the expression of a gene might vary across tissues, disease states and species without having to download any data.

Curators will have to adjust to the ways that data are changing, says Tanya Barrett, coordinator at GEO. A growing proportion of the data finding their way into repositories are derived from RNA sequences, which poses challenges: the files are larger, methods are still in flux and integration with conventional microarray data is difficult. But the biggest factor to limit data reuse could be cultural. Many researchers are reluctant to use data that are in different formats, or from other experimental designs or materials, says Ann Zimmerman, who studies data reuse at the University of Michigan in Ann Arbor. Familiarity could help to solve the problem, says Barrett. The more examples of data reuse that scientists see, the more ways they will find to reuse data. ■

NATURE.COM
To read a *Nature* supplement on genomics, see:
go.nature.com/ftkwlr

HEALTH

Wary approval for drug to prevent HIV

US regulators seek to mitigate risks of combined pill.

BY AMY MAXMEN

US regulators took a step into the unknown this week when they approved the first drug to prevent HIV infection. US Food and Drug Administration (FDA) commissioner Margaret Hamburg hailed the pill, Truvada, as a tool for reducing the rate of infection in the United States, where 50,000 people are diagnosed each year. But the drug combines low doses of two antiretroviral agents normally used to treat infection, and some researchers fear that its use in healthy people could have unacceptable side effects and spark the emergence of resistant viruses.

US insurers must now decide whether they will pay for Truvada, which costs roughly US$10,000 for a year's supply. Moreover, health-policy experts must script guidelines on how to prescribe it, and how to monitor side effects and HIV infections in people using the drug. "There are a lot of questions about how to implement it," says Connie Celum, an HIV researcher at the University of Washington in Seattle, who led a large trial[1] of the drug in East Africa and has begun studies to answer practical delivery questions, such as which subsets of people are at highest risk.

Developed by Gilead Sciences in Foster City, California, Truvada proved particularly effective in the East African trial[1], published last week: it reduced the incidence of HIV by 75% in people with partners who had been infected. In an earlier trial[2] in the United States, HIV incidence dropped by 44% in men who have sex with men.

But concerns emerged on 10 May at a public meeting of a panel that advised the FDA on its decision. Most members voted in favour of approval, but the researchers, doctors and patient advocates in attendance wrestled with the issue of drug resistance. The two drugs in Truvada, emtricitabine and tenofovir, are effective antiretroviral treatments, but trials have shown that viruses exposed to lower doses in the acute phase of infection can become resistant, said meeting attendees. In six people who tested negative on enrolment but turned out to be HIV-positive, the drugs were no longer effective. Another fear, unconfirmed in trials, was that people might not take the pill consistently, and might contract a strain of HIV that became drug-resistant as a result of exposure to low levels of antiretrovirals.

To mitigate these risks, the FDA requires that Truvada be prescribed only once an individual has tested negative for HIV. The agency also advises that people use the drug in combination with safe sex practices, and get tested for the virus every three months while taking it. Some experts at the advisory meeting proposed stricter policies, such as making the tests mandatory, but these were dismissed as impractical. Another idea was to limit the drug to specific populations who are at the very highest risk, such as homosexual people who use intravenous drugs, but the FDA adopted a vaguer category encapsulating anyone at high risk of contracting HIV. "We want to reach marginalized populations," says Celum, "and restricting access would mean that Truvada would be less likely to have a public-health impact."

Wayne Chen, acting chief of medicine at the AIDS Health Foundation in Los Angeles, California, regrets the decision to approve the drug, saying that condoms are cheaper and can be a more effective preventative. "The best thing would be to have this drug withdrawn from the market, and if it's not, there should at least be mandatory testing because we know that people don't take this as prescribed," he says, citing a Truvada clinical trial[3] in Africa that was ended prematurely because the drug was not preventing infection. Blood tests later confirmed that fewer than 40% of the study participants on Truvada had been taking the pills daily.

"Truvada is now the only technology we have that empowers women."

To proponents, however, the promise of the drug is bright. Salim Abdool Karim, director of the Center for the AIDS Programme of Research in South Africa in Durban, hopes that Truvada might soon be available in his country, where up to one-quarter of women have HIV by the age of 20. "Truvada is now the only technology we have that empowers women," he says. "I don't think we'll be able to slow the HIV epidemic in South Africa without something to protect them." ■

1. Baeten, J. M. *et al. N. Engl. J. Med.* http://dx.doi.org/10.1056/NEJMoa1108524 (2012).
2. Grant, R. M. *et al. N. Engl. J. Med.* **363,** 2587–2599 (2010).
3. Van Damme, L. *et al. N. Engl. J. Med.* http://dx.doi.org/10.1056/NEJMoa1202614 (2012).

A. NANTEL/SHUTTERSTOCK

valuable," agrees Enrico Petretto, a genomicist at Imperial College London. "We would never be in a position to look across multiple tissues and species with the money we have." But he cautions that using other people's data can be tricky. If data sets give contradictory outcomes, it is unclear whether that is because the underlying data contradict each other or because something went wrong with the analysis. "That's why people sometimes don't trust this," he says.

CHANGE OF PRACTICE

Still, few researchers are using the data to their greatest potential, says Alvis Brazma, a bioinformatician at the EBI. "Being able to reuse functional genomics data is a really new thing," he says. Researchers rarely download more than half a dozen data sets, and most use the data only to compare with their own results. Studies that use only other scientists' data to come up with new findings are still unusual.

That makes Butte and Khatri trailblazers. Another pioneer is Gustavo Stolovitzky, a computational biologist at the IBM Thomas J. Watson Research Center in Yorktown Heights, New York, who has used publicly available data to train algorithms to recognize gene signatures for diseases such as lung cancer, chronic obstructive pulmonary disease (COPD) and psoriasis. Not only can the algorithms distinguish lung cancer from COPD, they can also tell squamous-cell carcinoma from adenocarcinoma. "There is enough info in existing databases to predict disease in samples that algorithms have never seen before," Stolovitzky says.

Other efforts promise to unleash even more power from the growing repositories. In 2009, for instance, curators of ArrayExpress used their database to create the Gene Expression Atlas, which allows researchers to look at how the expression of a gene might vary across tissues, disease states and species without having to download any data.

Curators will have to adjust to the ways that data are changing, says Tanya Barrett, coordinator at GEO. A growing proportion of the data finding their way into repositories are derived from RNA sequences, which poses challenges: the files are larger, methods are still in flux and integration with conventional microarray data is difficult. But the biggest factor to limit data reuse could be cultural. Many researchers are reluctant to use data that are in different formats, or from other experimental designs or materials, says Ann Zimmerman, who studies data reuse at the University of Michigan in Ann Arbor. Familiarity could help to solve the problem, says Barrett. The more examples of data reuse that scientists see, the more ways they will find to reuse data. ■

NATURE.COM
To read a *Nature* supplement on genomics, see:
go.nature.com/ftkwlr

HEALTH

Wary approval for drug to prevent HIV

US regulators seek to mitigate risks of combined pill.

BY AMY MAXMEN

US regulators took a step into the unknown this week when they approved the first drug to prevent HIV infection. US Food and Drug Administration (FDA) commissioner Margaret Hamburg hailed the pill, Truvada, as a tool for reducing the rate of infection in the United States, where 50,000 people are diagnosed each year. But the drug combines low doses of two antiretroviral agents normally used to treat infection, and some researchers fear that its use in healthy people could have unacceptable side effects and spark the emergence of resistant viruses.

US insurers must now decide whether they will pay for Truvada, which costs roughly US$10,000 for a year's supply. Moreover, health-policy experts must script guidelines on how to prescribe it, and how to monitor side effects and HIV infections in people using the drug. "There are a lot of questions about how to implement it," says Connie Celum, an HIV researcher at the University of Washington in Seattle, who led a large trial[1] of the drug in East Africa and has begun studies to answer practical delivery questions, such as which subsets of people are at highest risk.

Developed by Gilead Sciences in Foster City, California, Truvada proved particularly effective in the East African trial[1], published last week: it reduced the incidence of HIV by 75% in people with partners who had been infected. In an earlier trial[2] in the United States, HIV incidence dropped by 44% in men who have sex with men.

But concerns emerged on 10 May at a public meeting of a panel that advised the FDA on its decision. Most members voted in favour of approval, but the researchers, doctors and patient advocates in attendance wrestled with the issue of drug resistance. The two drugs in Truvada, emtricitabine and tenofovir, are effective antiretroviral treatments, but trials have shown that viruses exposed to lower doses in the acute phase of infection can become resistant, said meeting attendees. In six people who tested negative on enrolment but turned out to be HIV-positive, the drugs were no longer effective. Another fear, unconfirmed in trials, was that people might not take the pill consistently, and might contract a strain of HIV that became drug-resistant as a result of exposure to low levels of antiretrovirals.

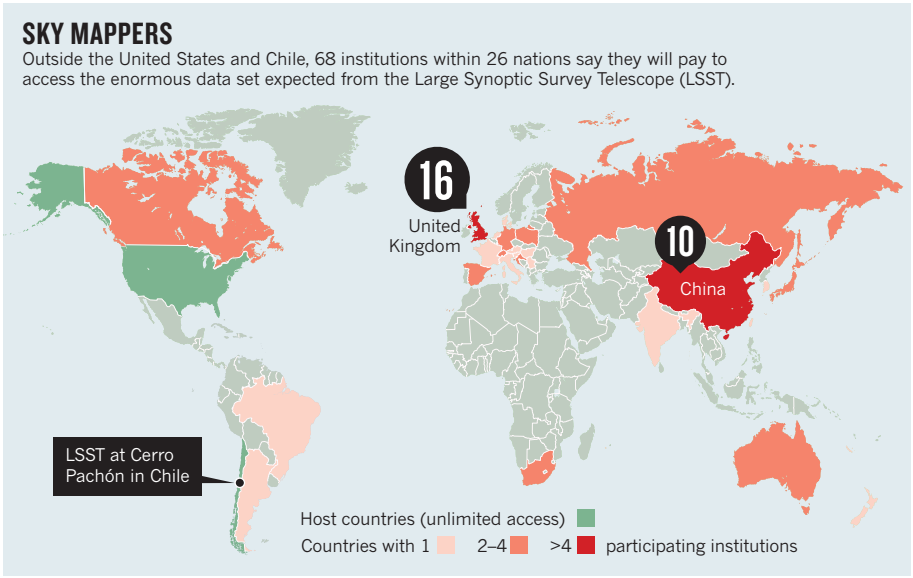
To mitigate these risks, the FDA requires that Truvada be prescribed only once an individual has tested negative for HIV. The agency also advises that people use the drug in combination with safe sex practices, and get tested for the virus every three months while taking it. Some experts at the advisory meeting proposed stricter policies, such as making the tests mandatory, but these were dismissed as impractical. Another idea was to limit the drug to specific populations who are at the very highest risk, such as homosexual people who use intravenous drugs, but the FDA adopted a vaguer category encapsulating anyone at high risk of contracting HIV. "We want to reach marginalized populations," says Celum, "and restricting access would mean that Truvada would be less likely to have a public-health impact."

Wayne Chen, acting chief of medicine at the AIDS Health Foundation in Los Angeles, California, regrets the decision to approve the drug, saying that condoms are cheaper and can be a more effective preventative. "The best thing would be to have this drug withdrawn from the market, and if it's not, there should at least be mandatory testing because we know that people don't take this as prescribed," he says, citing a Truvada clinical trial[3] in Africa that was ended prematurely because the drug was not preventing infection. Blood tests later confirmed that fewer than 40% of the study participants on Truvada had been taking the pills daily.

"Truvada is now the only technology we have that empowers women."

To proponents, however, the promise of the drug is bright. Salim Abdool Karim, director of the Center for the AIDS Programme of Research in South Africa in Durban, hopes that Truvada might soon be available in his country, where up to one-quarter of women have HIV by the age of 20. "Truvada is now the only technology we have that empowers women," he says. "I don't think we'll be able to slow the HIV epidemic in South Africa without something to protect them." ■

1. Baeten, J. M. *et al. N. Engl. J. Med.* http://dx.doi.org/10.1056/NEJMoa1108524 (2012).
2. Grant, R. M. *et al. N. Engl. J. Med.* **363,** 2587–2599 (2010).
3. Van Damme, L. *et al. N. Engl. J. Med.* http://dx.doi.org/10.1056/NEJMoa1202614 (2012).



SOURCE: LSST CORP.

ASTROPHYSICS

Cosmic survey finds global appeal

Partners line up to join the Large Synoptic Survey Telescope.

BY ERIC HAND

The past few years have not been the best of times for building observatories. But in a world of budget constraints and schedule delays, the Large Synoptic Survey Telescope (LSST) is bucking the trend. The US-led project to build the world's most powerful sky-mapping machine has nailed down international partnerships to fund project operations, which are intended to start in 2022. The commitments could help it to secure a final blessing from a key group: the board of the US National Science Foundation (NSF), which met this week in Washington DC.

From its perch atop Cerro Pachón in Chile, the proposed 8.36-metre telescope would map the entire southern sky every three nights, generating a wealth of data on transient events such as supernovae and passing asteroids, and helping to discern the nature of dark energy, which is accelerating the expansion of the Universe. Massive computing centres would store the data and allow astronomers around the world to access them remotely. Snapshots of portions of the sky would be released every minute, and more detailed maps would come out once a year.

The project reflects a shift in astronomy from the study of individual objects to surveys and big data. It has wide appeal: in 2010, it came out top of a decadal survey of US funding priorities in astronomy and astrophysics. The telescope is expected to produce many more data than the Sloan Digital Sky Survey, a highly productive northern survey. In less than two nights, the LSST will cover the same amount of sky as the Sloan managed in 8 years.

Organizers are confident that they will secure construction money. Aided by a total of US$30 million from philanthropists Bill Gates and Charles Simonyi, the project has already cast its primary mirror. The US Department of Energy (DOE) has committed $160 million towards a 3.2-gigapixel camera, and the NSF expects to be able to provide $466 million to build the rest of the telescope.

T. MASON/MASON PRODUCTIONS/LSST CORP.

The Large Synoptic Survey Telescope (artist's impression) will map the sky every three nights.

But the foundation is concerned about the high cost of operating the data centres that will deal with the telescope's output of 13 terabytes of data per night. Anthony Tyson, a physicist at the University of California, Davis, and director of the LSST project, says that in 2011, the NSF told him to shift the emphasis of his international fund-raising efforts from construction to operations.

The project is pioneering an innovative partnership model. In most astronomical consortia, members get a share of the telescope time that is proportional to the money they have put in. But with the LSST, institutions buy access to data: $20,000 in annual support secures access for a principal investigator, two postdoctoral researchers and unlimited graduate students. "It's a good deal, right?" says Sidney Wolff, president of the non-profit LSST Corporation in Tucson, Arizona.

Tyson found that recruiting partners was easy. He says that word would get out among astronomers in a country, and multiple institutions would soon be asking to join. "It mushroomed," he says. "It was limited purely by the number of hours I could stay awake." By the end of April this year, he had met his goal: 68 letters of intent from institutions across 26 nations, enough to cover nearly one-third of the annual operations costs of $37 million (see 'Sky mappers'). However, the first round of fund-raising has been closed to new partners, and Tyson says that some astronomers in countries such as France are disappointed that they missed out. (Astronomers in the United States and the project's host country, Chile, will have free, unlimited access to the data.)

The international support has reassured the NSF: as *Nature* went to press, the board was expected to approve the project on 18 July. Approval would allow the NSF to ask Congress for construction funding from 2014. "We're fairly confident," says Steven Kahn, a physicist at SLAC National Accelerator Laboratory in Menlo Park, California, and deputy director of the LSST project. "We've had lots of hurdles put in our path, and we've jumped over them."

But the project could still be endangered if the NSF and the DOE don't get along. In 2010, the foundation pulled out of a plan to build an underground laboratory for DOE experiments in South Dakota. "There's still a lot of nervousness about interagency collaboration," says David MacFarlane, an astrophysicist at SLAC and chairman of the board of the LSST Corporation. But the agencies have drawn up a formal agreement that could help to reassure the NSF board that the collaboration is on solid ground.

Andy Woodsworth, a physicist in Victoria, Canada, who at the end of May chaired an external review of the project, says that the LSST has already found its footing. "The time has come for this sort of survey," he says. ■

PUBLISHING

Europe joins UK open-access bid

Britain plans to dip in to research funding to pay for results to be freely available.

BY RICHARD VAN NOORDEN

Being the first to try something new is nerve-wracking — so it is always a relief to see someone else follow your lead. When the UK government announced on 16 July that it would require much of the country's taxpayer-funded research to be open-access from April 2013, it was not immediately clear whether the move would set a trend or prove to be an isolated gamble — one that would leave the United Kingdom essentially giving away its research for free while still paying to read everyone else's.

But the next day, the European Commission (EC) matched the United Kingdom's vision, launching a similar proposal to open up all the work funded by its Horizon 2020 research programme, set to run in the European Union (EU) from 2014 to 2020 and disburse €80 billion (US$98.3 billion). The details will be negotiated over the next year, but EC vice-president Neelie Kroes emphasized the momentum that open access has already acquired. "We are leading by example, making EU-funded research open to all — and we are urging member states to do likewise, so that sooner, rather than later, all nationally funded research will follow." The EC says that it is aiming for 60% of all European publicly funded research articles to be open access by 2016.

The announcements weren't unexpected. Britain's policy follows last month's government-commissioned Finch report on open access (see *Nature* **486,** 302–303; 2012), itself the culmination of more than a year of debate. The EC has made no secret of its support for open access, having run a pilot trial that covers some 20% of the budget of its current research-funding scheme, the Seventh Framework programme.

But coming in such quick succession, the statements mark Britain and Europe's determined plunge into an uncertain open-access transition that will dramatically shift the incentives for scientists, journal publishers and research institutions over the next five years.

Other funding bodies such as the US National Institutes of Health (NIH) and Australia's National Health and Medical Research Council already mandate a degree of open access. These agencies compel researchers to make their work publicly available in a separate repository within 12 months of publication — a version of 'green' open access that coexists with conventional subscription-based publishing.

But the UK Finch report advocated that authors should make their work free to read immediately on publication by paying publishers up front — the 'gold' open-access model. This is controversial among some researchers who argue that it sustains publishers' already high profits by eating into funds that could be used for research, and that the Finch report has played down the value of green repositories.

Although the UK policy recommends the gold route, it includes a much larger role for green open access than the Finch report envisaged. The plan is set out by Research Councils UK (RCUK), the umbrella body for the nation's seven research councils that award government grants. To cover the up-front charges for gold papers, the RCUK will pay 1–1.5% of its £2.8-billion annual research budget in block grants to research institutions. Each will use the money to set up a publications fund to pay for its researchers' papers, with the size of the award being proportional to each institution's research activity in recent years. Prepaid gold papers must have a liberal publishing licence, making text and data free to mine or reuse, the RCUK policy adds.

For journals that don't offer gold open access, the RCUK insists that they allow authors to deposit the final peer-reviewed version of a paper online within 6 months of publication (a system with which *Nature* complies). A longer embargo of 12 months is allowed for the arts, humanities and social sciences. The RCUK says that journals that don't allow either route should be shunned by researchers. The EC proposal matches this mixed green–gold model, right down to the 6- and 12-month publishing embargoes, but allows individual researchers to pay any author fees from their own grants.

To enforce its policy, the RCUK will probably tie compliance to future funding — much like the rule that the Wellcome Trust, a private UK research charity, announced in late June to beef up the 55% compliance of its own green–gold open-access mandate. The RCUK hopes after "a number of years" to approach the 75% compliance that the NIH has achieved for its green open-access policy, according to Astrid Wissenburg, chairwoman of the RCUK Impact Group, which is charged with increasing the economic and societal benefits of research-council funding.

"The fear that the UK ends up isolated is not going to happen."

If researchers do fall in line, the wide adoption of open access will shift everyone's publishing behaviours. Scientists may start discussing with universities where, and how much, they can afford to publish. Publishers and learned societies that rely on profits from library subscriptions will have to be more transparent about the costs of publishing. The latest open-access journals, such as *PeerJ* and *eLife*, may gain from the resulting melee (see *Nature* **486,** 166; 2012).

A large-scale change will depend on other countries following the United Kingdom and the EC; as *Nature* went to press, rumours were circulating that the US National Science Foundation was set to announce a new open-access policy of its own.

UK science minister David Willetts told *Nature*: "The fear that the UK ends up isolated is not going to happen — our policy will shape the international debate." ■

MORE ONLINE

TOP STORY

D. TIPLING/GETTY

Amazon's extinction debt still to be paid go.nature.com/rmo2d3

MORE NEWS

- Japan and Vietnam join forces to exploit rare-earth elements go.nature.com/cebfoj
- Tibetan glaciers shrinking rapidly go.nature.com/q3tsjc
- The environmental factors behind dolphin deaths go.nature.com/qmsmwi

VIDEO

V. M. ORTEGA

Raindrops falling on its head doesn't bother a hummingbird go.nature.com/gnqmdo

A. MCKEMEY

The release of transgenic mosquitoes is welcome news in Brazil, but less so in Key West.

PUBLIC HEALTH

Florida abuzz over mosquito plan

Biotech firm's bid to control dengue fever using genetically modified insects faces growing public opposition.

BY AMY MAXMEN

It took a decade for the biotechnology firm Oxitec to develop genetically modified mosquitoes whose progeny die before they can spread dengue fever. But it took only three months for Mila de Mier to gather 100,000 names from people opposed to the release of the mosquitoes in Key West, Florida, where the potentially lethal disease is making a comeback.

The US Food and Drug Administration (FDA) is currently reviewing an application from Oxitec, based in Abingdon, UK, and says that the mosquitoes will not be released without federal approval, which is not expected to happen soon. But that has not quelled the public furore, a sign of the communications challenge faced by those hoping to deploy genetically engineered (GE) animals for food or public health. An opaque system for reviewing applications does little to clear up the confusion.

"The more questions we ask, the more confused we are," says de Mier, a Key West business woman, who started the petition in April. "I started thinking, 'Oh my goodness, what if these mosquitoes bite my boys or my dogs? What will they do to the ecosystem?'."

The Oxitec mosquitoes are an engineered version of *Aedes aegypti*, the main transmitter of dengue fever. The modified males carry a lethal gene that is kept in check only by a special diet. They survive to mate with wild females, but the offspring die. In field tests conducted in Juazeiro, Brazil, the engineered insects shrank the *A. aegypti* population in an 11-hectare area by 85% over one year.

The online petition de Mier initiated to stop Oxitec has garnered national attention. It alludes to concerns that GE mosquitoes might harm people and that mosquito-eating native Florida species, such as bats, could go hungry. It also raises the prospect of other unintended consequences, such as the emergence of a deadlier dengue virus that gets around the absence of *A. aegypti*. In response, the company says that the virus already evolves in humans to optimize its fitness. It also notes that male mosquitoes do not bite, and that although a few engineered females might be released, any DNA they might transmit is not toxic or allergenic. Entomologists say that no animals in Florida feast solely on this species of mosquito.

Florida's negative reaction contrasts with that in Bahia state, Brazil, where residents in Juazeiro cheered the opening of an Oxitec mosquito-production facility on 7 July. Some Brazilians initially voiced concerns similar to those of de Mier and others, says Margareth Capurro, a biologist at the University of São Paulo, who led the Juazeiro trial. But she and her team engaged the community through meetings, radio and local television before seeking approval for their trial from Brazil's agency for biotechnological safety, CTNBio. Capurro continues to spread the message that GE mosquitoes are not a threat and that they fight a disease that residents know and fear. "We release the mosquitoes around 8 a.m., and the kids like to follow us," she says. "Sometimes you see them running back to older people in the village to explain what we're doing."

Dengue fever is a smaller problem in the United States than in Brazil, but health officials were alarmed when it reappeared in Florida three years ago after an absence of more than 70 years. Since 2009, 94 cases have been reported in Key West, and dengue prevention has become a top priority. Tourists often visit the area after stopping in dengue-infested countries, and a population of *A. aegypti* is there ready to spread the disease once it arrives, says entomologist Michael Doyle, director of the Florida Keys Mosquito Control District (FKMCD) in Stock Island, a taxpayer-funded operation that spends more than US$1 million a year to control *A. aegypti* in Key West with insecticides.

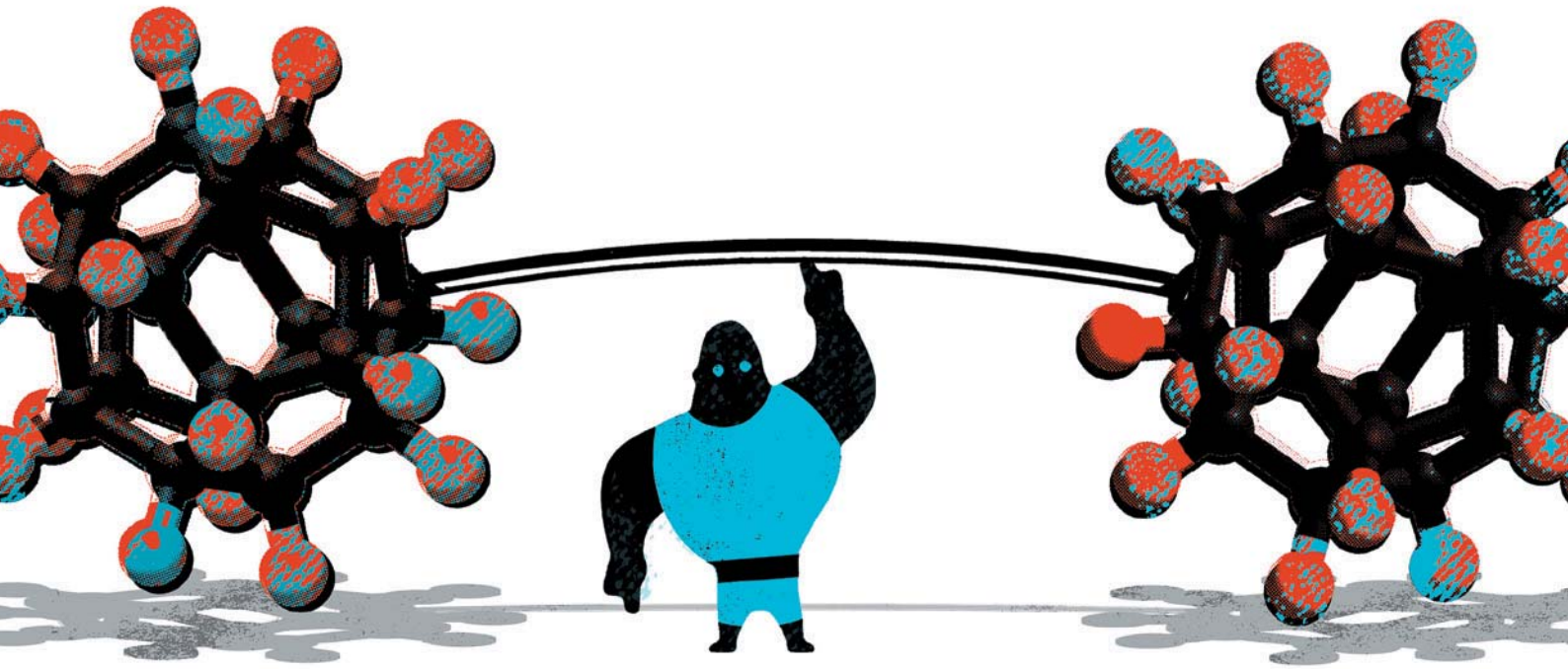
In 2010, the FKMCD asked Oxitec if it would do a field trial with its mosquitoes in Key West. Oxitec responded in its usual way. "We say we're delighted to help, and then we ask about the regulatory system in that country," says Hadyn Parry, the firm's chief executive.

In November 2011, Oxitec applied for approval by the FDA, which regulates GE animals that affect other animals (those that affect plants, such as crop pests, are regulated by the US Department of Agriculture). Although the FDA process is neither clear-cut nor rapid, a media report that month prompted concerns among residents after it suggested that officials were hoping for a mosquito release as early as January 2012. Doyle then organized a public meeting in March that de Mier attended, and which prompted her to start her petition. "I thought that if I presented the facts in a reasonable manner, people would respond in a reasonable way. But that's not happening," Doyle says.

Parry has offered to speak to de Mier about her concerns, but she has declined. Any rapprochement is unlikely to accelerate what is expected to be a long journey through the FDA's regulatory pipeline. ■

CORRECTION

The News story 'Palm-oil boom raises conservation concerns' (*Nature* **487,** 14; 2012) wrongly attributed 27% of Indonesia's deforestation to palm-oil planting. This figure was only for the area of Ketapang. It also said that this figure would rise to 40% by 2020, but that is the area in Ketapang projected to be given over to oil palm and is not due solely to deforestation.



SUPERHUMAN ATHLETES

Enhancements such as doping are illegal in sport — but if all restrictions were lifted, science could push human performance to new extremes.

BY HELEN THOMPSON

UK sprinter Dwain Chambers faces the race of his life next month, as he attempts to win an Olympic medal at the 2012 games in London — and complete a long journey back from the disgrace of his 2003 suspension for doping.

Chambers, who has devoted much of his time since then to persuading others to steer clear of performance-enhancing drugs, has admitted to using six different substances banned by the sporting authorities. These included two anabolic steroids — a designer drug and a testosterone cream — to accelerate recovery; the hormone erythropoietin (EPO), which increases production of red blood cells, to allow him to do more repetitions in training; human growth hormone for recovery; a thyroid hormone called liothyronine to decrease sluggishness; and a narcolepsy drug called modafinil to increase mental alertness and reaction time.

The quest for ultimate enhancement is as old as the games: the Greek physician Galen passed on knowledge from the ancient games to the Romans, praising the effects of eating herbs, mushrooms and testicles. But Chambers' story is just one example of how today's competitors are taking that quest to a whole new level.

ILLUSTRATIONS BY GARY NEILL

LET THE GAMES BEGIN
Science and the Olympics
nature.com/olympics2012

"There's an arms-race quality to performance-enhancing technologies in sport," says Thomas Murray, former president of the Hastings Center, a bioethics and public-policy foundation in Garrison, New York.

An amateur cyclist, Murray is among the many sports fans appalled by the seemingly endless string of doping scandals that result. "I could probably do a four-mile climb much better with EPO," he says, "but I could also do it much better if I put a motor on my bike." That's not the point of sport, he says, and neither are drugs — an attitude shared by the International Olympic Committee and just about every other professional and amateur sports organization.

But others argue that enhancers have become so prevalent that the only realistic option is for the sporting authorities to let athletes use what they want, as long as they do it safely.

"If the goal is to protect health, then medically supervised doping is likely to be a better route," says Andy Miah, a bioethicist at the University of the West of Scotland in Ayr. "Better yet, the world of sport should complement the World Anti-Doping Agency with a World Pro-Doping Agency, the goal of which is to invest in safer forms of enhancement."

Science alone cannot resolve the ethical conundrum presented by this debate. But it can shed light on the purely technical question: if performance-enhancing techniques were allowed, how far could the human body go?

POWER PILLS

For strength and power, the best-known drugs are probably those in the vast family of anabolic steroids, a group that is constantly expanding as the structures get slight modifications in a bid to evade detection in drug tests. "There are about 2,000 different tweaks you could do to a steroid molecule that would all probably make you big and strong," says Don Catlin, a pharmacologist at the University of California, Los Angeles. The compounds mimic the way testosterone works in the body, triggering protein synthesis and building more muscle tissue. A course of steroids combined with exercise can translate to a 38% increase in strength in men, potentially more in women.

Another popular strength enhancer is human growth hormone, which increases levels of the protein insulin-like growth factor 1 (IGF1). This spurs muscle growth, although it is debatable whether or not that growth actually does increase strength. In the only study to show positive effects in recreational athletes[1], those taking human growth hormone saw their sprinting capacity increase by 4%. That may seem small, but it could make all the difference for, say, a 50-metres freestyle swimmer or a 100-metres sprinter, says Kenneth Ho, an endocrinologist at the University of Queensland in St Lucia, Australia, who co-authored the study. "If you look at what breaks records, it comes down to 0.01 of a second."

In endurance sports, in which strength is less important than increased stamina, athletes can get dramatic results from blood doping, which aims to increase the number of oxygen-carrying red blood cells. They can accomplish this through blood-cell transfusions or by taking EPO. In one study[2], blood doping increased normal humans' stamina by 34%, and in another[3], it allowed them to run 8 kilometres on a treadmill 44 seconds faster than they could before. And work published last month[4] by Max Gassmann and his colleagues at the University of Zurich in Switzerland, there are signs that the hormone has an effect on the brain, increasing an athlete's motivation to train.

Drugs currently in the pipeline at pharmaceutical companies may also find themselves being co-opted for illicit use by athletes. One family, designed to treat muscular dystrophy and other muscle-wasting disorders, inhibits the activity of myostatin, a protein that keeps muscle growth under control. Similarly, a group of drugs called HIF stabilizers, which are aimed at treating anaemia and kidney disease, regulates a protein that turns on genes for the production of red blood cells, including the gene for EPO. And there may be a part for cognitive enhancers to play, too. "There's a range of compounds coming out that try to improve the ability to think more clearly when you're fatigued," says Chris Cooper, a biochemist at the University of Essex in Colchester, UK.

Improvements don't just come from the pharmacy. Athletes also rely heavily on nutritional supplements, which are legal. "They're 98.5% hype," says Conrad Earnest, an exercise physiologist at the University of Bath, UK. But one supplement that does work for some athletes is creatine, which contributes to the synthesis of the energy carrier molecule ATP during exercise. Earnest estimates that athletes taking creatine could see their performance improve by as much as 8%.

Another effective supplement is beetroot juice. Researchers at the University of Essex have found that the nitrate present in the juice increases nitric oxide levels in the body, allowing muscles to use oxygen more efficiently. As a result, the team found that divers could hold their breath for 11% longer than normal[5], which could help swimmers who want to minimize the number of breaths they take in short-distance events.

Most of these performance enhancements come with a slew of side effects, however. Steroids can cause high blood pressure, thickening of the heart valves, decreased fertility and libido, and changes such as chest hair in women and shrunken testicles in men. And boosting the number of red blood cells thickens the blood, increasing the risk of having a stroke.

Adding to the uncertainty, a number of the drugs are used to treat serious diseases such as cancer, AIDS and muscular dystrophy, so they have been tested largely on desperately ill patients with below-normal levels of growth factors and hormones. It is hard to know how to extrapolate those data to the sports arena, says Cooper. "Elite athletes are very different beasts from normal people in the sense that they're genetically enhanced," he says, "because they've been selected to be good at what they're doing and they have a lot of training."

Furthermore, testing in healthy people — subjecting them to the dosages and combinations that athletes are likely to take — would be an ethical can of worms. Because of that, says Charles Yesalis, an emeritus professor of sports science at Pennsylvania State University in State College, "there's no way to know what advantages different combinations of steroids, nutritional supplements and specialized diets could produce. It's a witches' cauldron."

CODE BREAKING

Gene doping — enhancing performance by adding or modifying genes — has been the subject of locker-room gossip for the past ten years. There are plenty of natural mutations for which to wish. The Finnish cross-country skier Eero Mäntyranta, who won three gold medals in the early 1960s, had a mutation that made his body's EPO receptors more efficient. In 2004, a toddler made headlines for having a mutation that disabled myostatin, giving him the physique of a petite body builder. And the gene that encodes angotensin-converting enzyme, which has been hailed as the gene for physical performance, has one variation known to boost endurance by increasing oxygen delivery capacity and capillary density, and another that is associated with muscle growth and strength[6,7].

Advances in gene therapy could one day make it possible for any athlete to enhance their DNA. For example, in experiments aimed at treating muscular dystrophy in the elderly, a group led by physiologist Lee Sweeney of the University of Pennsylvania in Philadelphia introduced a gene to cause over-expression of IGF1 in mice. The treatment boosted muscle strength of young adult mice by 14%, earning the rodents the nickname 'mighty mice'[8].

Other researchers are turning genes on and off with drugs. In 2008, Ronald Evans and his colleagues at the Salk Institute for Biological Studies in La Jolla, California, worked with GW1516, a drug that activates a gene that increases the ratio of 'slow-twitch' to 'fast-twitch' fibres in muscle. As the names suggest, slow-twitch fibres contract more slowly than fast-twitch, but they are more efficient at aerobic

activity. Evans and his team found[9] that in mice, GW1516 combined with exercise increased the rodents' endurance by 70%.

However, both Evans and Sweeney are sceptical about how useful athletes will find such therapies. "In humans, I expect the same general relationship — the under-exercised will be the ones who will have the most benefit from exercise mimetics," says Evans. "My view is that endurance athletes are physically advantaged and will have the least benefits."

Gene therapy has its share of health risks, including potentially severe immune reactions to the viruses used to ferry genetic material into cells. The results may also be hard to control. "If you're going to turn a gene for something like EPO on, you better be able to turn it off," warns Catlin. Gene doping, he says, "is not a good idea, but I wouldn't be surprised if someone's out there trying it".

"WHAT WE'LL SEE IS THE EMERGENCE OF ALL KINDS OF NEW SPORTS."

HUMAN 2.0

Drugs are not the only way to potentially enhance performance. Surgery and, ultimately, technological augmentations could also help athletes towards the podium. Baseball pitchers who have undergone surgery to replace a damaged elbow ligament with tissue from a hamstring or forearm tendon claim that they can throw harder after the two-year rehabilitation process. But Scott Rodeo, an orthopaedic surgeon at the Hospital for Special Surgery in New York City, warns that the science doesn't back up the stories. "To truly say you're making this elbow better would be a bit of a stretch," says Rodeo.

Replacing entire joints would be unlikely to work for an elite athlete: too many screws could come loose and the artificial joint wouldn't quite match the mechanics of a natural one. The materials would also wear out within a few years under the physical demands of elite sport. Still, Rodeo says, that assessment could change if researchers make major advances in engineering skin, tendons and other replacement body parts in the laboratory.

Miah sees potential in more imaginative surgical enhancement. "Consider using skin grafts to increase webbing between fingers and toes to improve swimming capacity," he says. "These kinds of tweaks to our biology are likely ways that people would try to gain an edge over others." Another frontier is nanotechnology, adds Miah. Researchers are already experimenting with blood supplements based on oxygen-carrying nanoparticles for use in emergency situations. From there, he says, "there is a lot of discussion about the possibility of biologically infused nanodevices that could perpetually maintain certain thresholds of performance".

Mechanical prosthetics are already a reality, such as the 'cheetah-style' legs used by amputees including Oscar Pistorius from South Africa, a Paralympic gold medallist who was approved this month to run in the 2012 Olympics. But scientists are split on whether current artificial limbs actually confer an advantage over the flesh and blood variety.

Bryce Dyer, a prosthetic engineer at the University of Bournemouth, UK, explains that although Pistorius's spring-like prosthetics allow him to speed up at the end of a race, they put him at a disadvantage coming out of the crouch at the start of a race or when turning a curve. "When he's running straight ahead, he eventually hits a natural state of harmony like bouncing on a trampoline," says Dyer, "but then he sometimes runs right off the track because he can't turn."

Pistorius's prosthetics lack the stiffness of a human ankle and can't generate the same forces as they hit the ground. To get around this, Pistorius pumps his legs faster. "It's a biomechanically distinct way of running fast, but there's no evidence that it's advantageous," says Hugh Herr, a biomechanical engineer at the Massachusetts Institute of Technology (MIT) in Cambridge.

Technology might get around these problems. "Stepping decades into the future, I think one day the field will produce a bionic limb that's so sophisticated that it truly emulates biological limb function. That technology will be the Olympic sanctioned limb," says Herr, whose lab at MIT is currently working on a bionic running leg. "Without any such human-like constraints, the Paralympics limb will become [the basis of] this human–machine sport like racecar driving."

According to Herr, performance-enhancing technologies will advance to a point at which they will not only extend human limits, they will demand an Olympics all of their own. "For each one there will be a new sport — power running, and power swimming, and power climbing," projects Herr. "Just like the invention of the bicycle led to the sport of cycling. What we'll see is the emergence of all kinds of new sports." ■

Helen Thompson *is an intern in* Nature*'s Washington DC office.*

1. Meinhardt, U. *et al. Ann. Intern. Med.* **152,** 568–577 (2010).
2. Buick, F. J., Gledhill, N., Froese, A. B., Spriet, L. & Meyers, E. C. *J. Appl. Physiol.* **48,** 636–642 (1980).
3. Williams, M. H., Wesseldine, S., Somma, T. & Schuster, R. *Med. Sci. Sports Exerc.* **13,** 169–175 (1981).
4. Schuler, B. *et al. FASEB J.* http://dx.doi.org/10.1096/fj.11-191197 (2012).
5. Engan, H. K., Jones, A. M., Ehrenberg, F. & Schagatay, E. *Respir. Physiol. Neurobiol.* **182,** 53–59 (2012).
6. Montgomery, H. E. *et al. Nature* **393,** 221–222 (1998).
7. Williams, A. G. *et al. Med. Sci. Sports Exerc.* **37,** 944–948 (2005).
8. Barton-Davis, E. R., Shoturma, D. I., Musaro, A., Rosenthal, N. & Sweeney, H. L. *Proc. Natl Acad. Sci. USA* **95,** 15603–15607 (1998).
9. Narkar, V. A. *et al. Cell* **134,** 405–415 (2008).

BY DANIEL CRESSEY AND EWEN CALLAWAY

TEAM SCIENCE

The Olympics is a vast experiment in human performance, sport technology and global travel. Nature *meets some of the scientists behind the scenes.*

Science has had a hand in every aspect of the Olympic and Paralympic Games. For the thousands of athletes, researchers have helped to develop training techniques, schedules, diet, equipment and doping checks. For the millions of spectators about to descend on London, they have contributed to urban planning, crowd control, public health and security. For the billions watching at home, they have shaped the technology that will measure athletic feats and beam them worldwide.

Yet those scientists toil in the background, understandably overshadowed by the sporting spectacle. Here, *Nature* profiles four scientists whose work will contribute to the giant human experiment that is the Olympic Games.

THE PSYCHOLOGIST

In 2000, the Spanish basketball team in the Paralympics learning-difficulties category swept the board to win all of their games and take gold. There was just one problem: many of the team were not intellectually disabled. After the scandal was revealed by an undercover journalist, the team was stripped of its medals, and anyone with a learning disability was excluded from the next two Paralympic Games.

This year, in London, they can return, in athletics, swimming and table tennis. Jan Burns, head of the Department of Applied Psychology at Canterbury Christ Church University, UK, is one of the key scientists responsible for ensuring that the athletes qualify for competition.

Intellectual disabilities are difficult to police because, unlike most physical disabilities, they are not always obvious. In the wake of the Spanish fiasco, the International Paralympic Committee and Inas, the international federation for para-athletes with an intellectual disability, sponsored an international research group to solidify the criteria for 'eligibility' (existence of a disability), and 'classification' (impairment of ability to play the sport).

Burns, a specialist in intellectual disability, joined the research group in 2009 and became head of 'eligibility' at Inas. "She had an incredible interest in the interaction and the application of these psychological concepts in this kind of an environment," says Peter Van de Vliet, the medical and scientific director of the International Paralympic Committee, based in Bonn, Germany. The criteria Burns helped to develop have paved the way for intellectually disabled athletes to return to the Paralympics.

According to the rules now, an athlete is eligible if he or she has had a developmental delay before the age of 18; has an IQ of no more than 75; and has a 'significant limitation' in adaptive behaviour such as social skills. Eligible athletes are then subjected to a battery of tests to show that they classify as disabled for a specific event. A swimmer, for example, would first be assessed on skills that are generically useful in sports, such as reaction time. Then his or her swimming performance would be compared with that of other athletes. Burns points at research showing that people with intellectual disabilities tend to take more strokes to cover a given distance, so classifiers will video swimmers in competition and assess their stroke ratio to see whether it falls within the 'bandwidth' of disabled swimmers. All this has to be comprehensively documented and reviewed by multiple researchers so that the system is robust against fraud.

Burns is currently working a hectic schedule juggling her Paralympics work, her regular academic job and huge interest from the world's media. "I'm currently going through and checking everybody's file, making sure we know enough about everybody who's come through the system," she says. During the Paralympics, which run from 29 August to 9 September, "I'll be around ensuring that the classification goes well and to be on hand if we do have any issues".

Work is already under way to see whether more sports can be added for the 2016 Paralympics in Rio de Janeiro, Brazil. This involves working out the skills that athletes need to play a sport, and how intellectual disabilities might affect performance — for example, pattern recognition might be relevant to the complex plays in some team sports. The early betting is that the Rio Paralympics will include rowing and will give a second chance to the game that started the story: basketball.

LET THE GAMES BEGIN
Science and the Olympics
nature.com/olympics2012

O. SCARFF/GETTY

Christiaan Bartlett will be part of a huge team manning a 24/7 drug-testing lab north of London.

THE DOPING DETECTOR

In this summer's 100-metre sprint, Usain 'Lightning' Bolt will attempt to hold onto his title of 'world's fastest man' against a younger and currently fleeter Yohan Blake. Yet one of the fiercest battles in the Olympic Games will play out in a giant, custom-built suburban laboratory in Harlow, 35 kilometres north of the Olympic village. Here, anti-doping experts will apply the most sophisticated tools in their molecular arsenal in the seemingly Sisyphean pursuit of those athletes who take performance-enhancing drugs.

The lab will screen for dozens of stimulants, steroids and other banned substances. Christiaan Bartlett, a senior scientist at the King's College London Drug Control Centre, which is running the Harlow lab, will direct the testing for biological drugs such as the blood-boosting hormone erythropoietin (EPO) and human growth hormone. Anti-doping science is notoriously — some say unnecessarily — secretive; Bartlett says that he cannot reveal what drug-detection techniques will be rolled out at the London games. All he will disclose is this: "We've got the most sophisticated equipment, we spent the past year or so developing and validating new techniques that will give us increased sensitivity in all of our areas."

The first challenge for Bartlett and his 150 or so colleagues lies in handling the sheer volume of urine and blood that Olympic and Paralympic athletes will be required to submit for testing during the games — collected from as many as 7,000 athletes days before and immediately after sporting events. Samples will arrive hourly, with one part prepared for testing and the other frozen as a back-up, and the lab will run around the clock to turn around most tests within a day. Bartlett has already decamped from his home in south London to live closer to the lab, and he knows that his weekends won't be spent watching sport.

The vast majority of the tests he oversees will come back with an all-clear. If the King's lab turns up any banned substances, scientists there will immediately inform the International Olympic Committee and other sport authorities, who will initiate an investigation and possibly disciplinary action.

Legitimate drugs are one target for Bartlett, who worked previously in food sciences and toxicology. Pharmaceutical companies such as Roche, Amgen and GlaxoSmithKline now routinely share information about drugs in their pipelines that could potentially be used by athletes. Months after the US Food and Drug Administration approved a new class of red-blood-cell boosters called CERAs in 2007, anti-doping scientists had developed a test for them. The test came too late for the Beijing games the following year, but retrospective testing stripped the men's 1,500-metre winner, Rashid Ramzi, of his gold.

Increasingly, dopers are turning to illegal labs in India, China and elsewhere that crank out drugs such as EPO that have been tweaked chemically to evade testing. Bartlett says that his team is ready. The test for EPO, for example, is designed to detect any forms of the protein made using genetic engineering, because these tend to be less acidic than the natural stuff.

Athletes participating in the London Olympiad will be the most heavily tested in the history of the games, but will that make them the cleanest? Bartlett is cautiously optimistic. Many countries screen their athletes before departing for London, and some sports have begun to use 'biological passports' that chart characteristics of athletes' blood over time, looking for changes that might signal illicit performance enhancement, even when a substance such as EPO cannot be found. "The general message is: athletes, if you're coming to London, beware," Bartlett says.

THE FLUID MODELLER

At the Beijing Olympics in 2008, athletes smashed 25 world records in swimming, more than in any other sport. Many gave the credit to high-tech swimsuits, which cut down on drag. But after Beijing, the international body that governs competitive swimming introduced rules that limited the advantage that could be gained from swimsuits, leaving athletes looking for other ways to gain an edge. British swimmers turned to fluid-dynamics researcher Stephen Turnock.

Turnock's speciality is hydrodynamics, particularly in ship design. It wasn't a huge leap to study how air or water flows around the human body, and for the past three years he has directed the Performance Sports Engineering Laboratory (PSEL) at the University of Southampton, UK. The lab previously worked with the British cycling team to devise more aerodynamic riding positions, which may have played at least a small part in the 14 medals that 'Team GB' cyclists brought home from Beijing. Swimming needed similar help, he says. "What British swimming lacked was an understanding of what the hydrodynamic forces were during the swimming processes."

Applying a scientific approach to swimming performance has proved a challenge, however. "With cyclists, you put someone in a wind tunnel and say 'What's your best position to lower your resistance?'," says Turnock. "It's a complex process to get the instrumentation right but that's a relatively simple bit of fluid dynamics because, typically, most of a position of the cyclist and the body is relatively fixed." But with swimming, a whole range of factors come into play, including roll along the length of the body, movement of the arms and the legs, forces that are transmitted from limbs to water and the effects of water pressure and movement in turn on the shape of the body. "There are so many variables and it's all happening quite quickly in a very noisy environment. It is very difficult to repeat exactly the same conditions every test run," says Turnock. "By the time you've got all that

London bound: researchers will use flight itineraries to predict the risk of disease outbreaks during the Olympics.

SOURCE: BIO.DIASPORA AT ST MICHAEL'S HOSPITAL, TORONTO

uncertainty in there it's quite challenging."

The team at the PSEL devised some technical solutions. Its main system is based around a portable winch, which pulls a swimmer through the water fractionally faster than they would normally swim, a technique called overspeeding. Working at pools in which top British swimmers train, Turnock's team measured the tension in the winch line to assess changes in water resistance, and the researchers videoed lap after lap to see how, for example, adjusting posture or even the position of a swimming cap might change water flow and speed. "We can examine what they've done pretty much as soon as the swimmer gets out of the pool," says Turnock, who says that the information is all fed back to coaches and athletes.

Turnock's team has also been tackling some broader questions about training. Using the winch system, a willing team member and a full-body wax, he explored how body hair affects resistance in the water. (Answer: smoother is faster.) The group is also using computer modelling of the musculoskeletal system to work out how to improve the efficiency of swimming strokes.

Turnock's work with Britain's swimmers wrapped up well before the start of the games. But he hopes that what he has learned about the hydrodynamics of human bodies might feed back into his work on marine systems, such as designing rudders that adopt a more hydrodynamic shape under water stresses. He also improved his own swimming and, he says, "I can shout learned things at my children when they learn to swim now."

THE DISEASE TRACKER

Kamran Khan will not be anywhere near London during the games. The medical researcher will be sitting thousands of kilometres away at the University of Toronto in Canada. But he will be watching.

Organizers have predicted that several million people, from all over the world, will descend on London — along with their viruses and bacteria. Khan is part of an international team that is testing strategies for predicting the spread of diseases, such as some potential new strain of flu, as the crowds arrive.

In all probability, they will see nothing; but it's the 'what if' that keeps Khan awake. Disease outbreaks have been associated with mass gatherings in the past, including a spike in measles around the 2010 winter Olympics in Vancouver, Canada, and an influenza outbreak linked to a Catholic youth festival in Sydney, Australia, in 2008. "With as big a mass gathering as the Olympic Games we want to think about the potential for health threats, particularly for infectious diseases, to move around the world," says Brian McCloskey, the UK Health Protection Agency's national lead for the Olympics.

To assess those threats, Khan will be using the Bio.Diaspora project, which he has been running since he set it up in 2008. This web-based computer program brings together information on billions of flight itineraries, allowing researchers to see how people are moving around the world. To gauge the risk of those people carrying a pathogen, it will link up with disease surveillance information collected in real time during the games, such as by the HealthMap project at Children's Hospital Boston in Massachusetts, which trawls through news, reports from health-care systems and social-media chatter for signs of emerging infectious threats. If, for example, a new form of flu emerges in China, Khan can piece together a picture of the disease's spread and use it to predict the likelihood of an outbreak reaching London. This type of early warning might give health officials crucial extra time to warn the public and take preventative action.

In fact, the Olympics will be a major test of the utility of Bio.Diaspora and global-health systems during a mass gathering, says McCloskey. "It could be that it doesn't add a lot of value or it could be that it's critically important," he says. "We don't know the answer until we've done the experiment."

Khan hopes that the data on the global flow of people can inform the growing field of 'mass-gathering medicine', the study of the public-health risks posed by religious rallies, music festivals and sports events, which are attracting more people than ever before and from more-remote places. He likens global transport to "a network of arteries around the world. There are people moving through those arteries, there's a sort of physiology. And that normal physiology is disrupted or changed by certain types of events," he says. "These events have potential implications for global-health security, and we need to understand them better." ■

Daniel Cressey *and* **Ewen Callaway** *are reporters for* Nature *in London.*

COMMENT

OLYMPICS Sitting about is bad for brains and bodies that evolved to run **p.295**

OLYMPICS Will future Olympics be genetically enhanced? **p.297**

EPIGENETICS Twin studies probe how environment and genes interact **p.298**

PUBLISHING Spat over 'green' and 'gold' routes to open access continues **p.302**

N. BERGER/POLARIS/EYEVINE

The 'Berlin Patient,' Timothy Brown, has been cured of HIV since 2007. His story has renewed interest in cure research.

Towards a cure for HIV

Steven G. Deeks and **Françoise Barré-Sinoussi** present an international research agenda to seek out a cure for AIDS.

One of the greatest achievements of modern medicine has been the development of combination antiretroviral (ARV) therapy for HIV. Today, fewer than half of the world's people who need treatment have access to therapy. A substantial and sustained increase in funding will be required to effectively treat the global population (see 'Cost of managing HIV'). And this life-saving therapy has limitations — medicines have side effects and must be taken daily, and HIV can develop resistance. Clearly, a new approach to tackling HIV is needed.

In 2007, an HIV-infected man in Berlin received a transplant of haematopoietic stem cells from a naturally HIV-resistant donor, and then he stopped HIV therapy[1]. He has now been free of readily detectable virus in the absence of therapy for more than five years. In other words, he is cured. His experience suggests that HIV infection might one day be curable.

One of the key priorities of the International AIDS Society (IAS) is to promote and facilitate the search for a safe, affordable and scalable cure. The multidisciplinary IAS Scientific Working Group on HIV Cure has developed a broad and ambitious set of priorities for cure research (see 'Priorities for HIV cure research'). Some of these research questions have been pursued for decades, but the focus has been mainly on improving therapy or developing vaccines. Unique perspectives on these old questions will almost certainly be needed for cure research to succeed.

NATURE.COM
Read more about the HIV/AIDS pandemic:
go.nature.com/xey8dc

Cure research is not completely new. Several high-profile approaches were attempted soon after the development of combination therapy[2]. But these attempts failed, and the field shifted towards optimizing therapy so that it could be taken indefinitely. Since ▶

then, only a few scientists have continued to do cure research, working without a clear source of funding and despite a widespread sense that a cure is not possible. Increasing the number of scientists involved in this effort will be a crucial first step.

ARV therapy cannot cure HIV mainly because the virus is able to integrate its DNA into the genomes of long-lived immune cells called memory CD4 T cells, preventing the immune system from recognizing and clearing these cells. Determining how this 'latency' works, where infected cells reside and how latent infection could be reversed are the focal points of most ongoing cure research. Such work is based on the assumption that if all infected memory CD4 T cells could be identified and forced to make virus while ARVs inhibit that virus from infecting new cells, then the host immune system could identify and kill all infected cells, thereby achieving a cure. But it is unclear whether the immune system has the capacity to kill infected cells, even if they could be forced to start making virus. It is also unclear whether other cell types harbour HIV during long-term therapy, or whether ARVs completely inhibit the virus.

It may not be necessary to completely eradicate the virus in the individual, however. In about 1% of people infected with HIV, the virus is naturally controlled, such that their risks of disease progression and transmission are minimal. Scientists have studied these 'elite controllers' in search of a path towards developing a vaccine. Elite controllers could also provide clues about how to control, if not eliminate, established infection.

Funding remains a problem. The US National Institutes of Health (NIH) devoted about US$56 million to cure-related work in 2011, although it plans to invest more in the future. The French National Agency for Research on AIDS and Viral Hepatitis provided more than €7 million (US$8.6 million) last year. Since 2006, the state-funded California Institute of Regenerative Medicine has spent more than $40 million on gene-therapy approaches that would, for example, make cells resistant to HIV infection. Foundations have also been supportive. The Foundation for AIDS Research, based in New York, contributed $4.1 million to the effort in 2011. Still, although a price tag for the IAS research priorities cannot be calculated, it is clear that more resources — perhaps hundreds of millions of dollars per year — will be needed to find a cure.

This extra funding should not be diverted from other high-priority research areas such as vaccine development, nor from life-saving and underfunded treatment programmes. The push should come from new funding sources, such as private foundations or governments that are fighting growing epidemics (such as China, India and Brazil).

It is also not enough to develop and fund a research agenda. The research infrastructure must be improved. The community must establish large, multinational and multidisciplinary collaborations specifically for cure research. The NIH Martin Delaney Collaboratory grants ($14 million a year for up to five years), which are focused on collaborative efforts to this end, are a first step.

Scientists also need more support for HIV research in animals. Although they can infect macaques with the simian version of HIV, many scientists cannot afford to treat animals with ARV therapy over long periods to replicate the physiology of a human, as it costs tens of thousands of dollars per year per animal. Resources will need to be made available for the development, validation and optimization of such non-human primate models, and for the development of small-animal models for more high-throughput studies.

The research community must also address the ethical issues that arise in HIV cure research, in which scientists must test new and potentially very toxic drugs in individuals who are receiving long-term ARV therapy. By definition, these people have access to therapy and are doing well, so the potential risks to them will need to weighed against the potential benefits for the larger community. Strong community support is needed to ensure that patients and their care-givers are fully engaged and informed about the risks and benefits of curative studies.

The barriers to curing HIV are real, and they may prove to be insurmountable. Scientists who have in the past spoken about the promise of a cure have experienced backlash when the cure did not materialize. It is the responsibility of organizations such as the IAS to encourage and enable research in this area, but to do so responsibly, without hype or false promises. ■

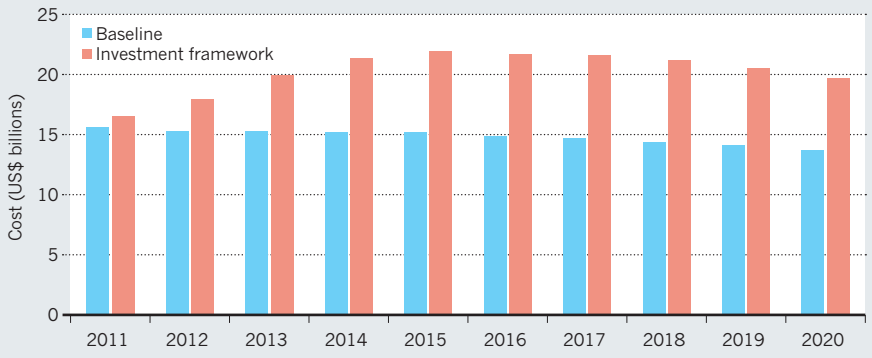
DEVISING A CURE

Priorities for HIV cure research

- Determine the mechanisms that maintain HIV persistence. This includes defining the role of latency, replication and cell proliferation.
- Determine the tissue and cellular sources of persistent HIV infection in animal models and in people who undergo long-term antiretroviral (ARV) treatment.
- Determine the origins of immune activation and inflammation in the presence of ARVs and their consequences for viral persistence.
- Determine host and immune mechanisms that control infection but allow viral persistence.
- Develop and validate assays to measure persistent HIV infection.
- Develop and test therapeutic strategies to safely eliminate latent infection, including reversal of latency and clearing of latently infected cells.
- Develop and test strategies to enhance the host response's capacity to control active viral replication.

COST OF MANAGING HIV

Researchers have proposed a plan (investment framework)[3] that would increase global access to HIV therapy and decrease the number of new infections in low- and middle-income countries over the coming decade. For the plan to succeed, however, a large increase over current (baseline) spending is needed.



SOURCE: REF. 3

Steven G. Deeks *is a professor of medicine at the University of California, San Francisco, California 94110, USA.*
Françoise Barré-Sinoussi *is a professor of virology at the Institut Pasteur, Institut National de la Santé et de la Recherche Médicale, Paris 75724, France. She is also IAS president-elect.*
e-mail: sdeeks@php.ucsf.edu

1. Hütter, G. *et al. N. Engl. J. Med.* **360,** 692–698 (2009).
2. The International AIDS Society Scientific Working Group on HIV Cure. *Nature Rev. Immunol.* (in the press).
3. Schwartländer, B. *et al. Lancet* **377,** 2031–2041 (2011).

S.G.D. declares competing financial interests; see go.nature.com/bjntyh.

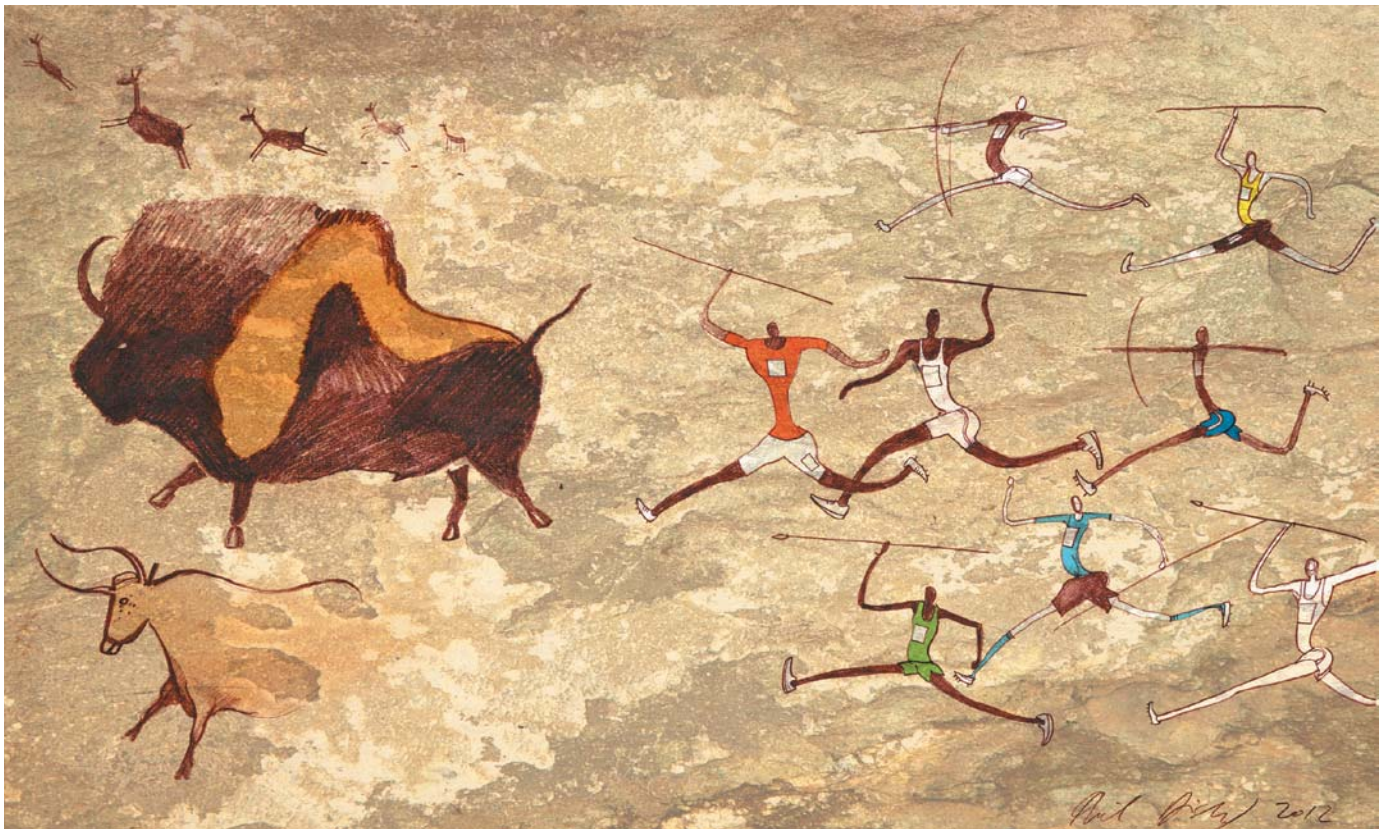


ILLUSTRATION BY PHIL DISLEY

Run for your life

Humans evolved to run. This helps to explain our athletic capacity and our susceptibility to modern diseases, argue **Timothy Noakes** and **Michael Spedding**.

The forthcoming Olympics in London will celebrate the performance capacity of humans and our remarkable ability to prepare our bodies and minds for specific tasks. But, at the same time as we are pushing our bodies to new limits in athleticism, we are experiencing unprecedented levels of relatively modern diseases such as obesity, diabetes and psychiatric and neurodegenerative disorders.

We, the authors, were both considering the modern paradox of elite athleticism and growing susceptibility to disease when we met at a sports conference in Glasgow, UK, in 2010. Noakes is a sports scientist who has run more than 70 marathons and ultramarathons. He was presenting data suggesting that humans' unmatched ability to dissipate heat when running, even when drinking sparingly, might have been a key element that enabled them to evolve from tree-living primates. Spedding, a pharmacologist presenting studies of how stress can increase the risk of psychiatric disorders, has run more than 100,000 kilometres and been a competitive athlete for more than 40 years. His brother, Charlie, holds the English marathon record and won Olympic bronze in 1984 by ignoring drink stations at crucial stages in the Los Angeles marathon. We began exchanging e-mails. Eventually, that correspondence coalesced into the theory we outline here.

Over millions of years, humans evolved from tree-dwelling apes to become *Homo sapiens*, capable of elite athleticism[1]. Simply put, we evolved to run. While early hominins were undergoing intense skeletal and metabolic changes, major changes also occurred in their brains[2]. We propose that these changes have rendered us dependent on mental and physical exercise to maintain brain health. Exercise doesn't just help muscles — it activates our brains, particularly through one pathway that helps to increase the number of neuronal connections.

Most humans today do not live in an environment where they must exercise regularly to chase down meat. For many, exercise is no longer an integral part of daily life, leading to a host of modern ailments.

LET THE GAMES BEGIN
Science and the Olympics
nature.com/olympics2012

"Most humans today do not live in an environment where they must chase down meat."

In short, we think that exercise is not just important for general health — it is essential to the molecular memory of who we are. Without it, we are at risk of being obese and diabetic, and of developing diseases linked to brain function, such as psychiatric disorders, dementia and even violent behaviour.

COOL AND SWEATY

The ancestors of modern humans were omnivorous apes with bodies that were more suited to living in trees than hunting in open habitats. Over the past few million years, the climate underwent dramatic shifts and Africa changed from a largely forested ecosystem to a more open savannah. Our ancestors, caught at the edge of the retreating forests, became less adapted for climbing

trees. By 2 million years ago, they had evolved a skeleton that could support walking and running — partly so that they could hunt by pursuing individual animals for hours at a time[1].

For more than 1 million years, there were no weapons other than stones or sharpened sticks. The best weapon was endurance. The predators had to outlast their prey, and so had many adaptations that enabled them to walk and run long distances, forcing their prey to gallop. Because four-legged animals cannot lose heat by panting and galloping at the same time, human hunters eventually drive their prey into heat stroke, so that the animal can be caught and killed with very simple weapons.

The ability of humans to dissipate heat comes from our lack of body hair and capacity to breathe through our mouths and to sweat at rates of up to 3 litres an hour, much more effective than panting. In a 3-hour hunt — or in a marathon — fit humans can safely lose up to 10% of their bodyweight[3].

At the same time, the hunter needed planning and spatial navigation to follow prey for hours, coupled with social interactions so that groups could work together to isolate prey[4]. Primitive hunter-gatherer societies have a uniquely human social structure with multiple interactions between non-family members, requiring advanced social skills[4]. Within a relatively short time, our ancestors' skeleton, brain, spatial tracking, communication and ability to dissipate heat shifted dramatically, allowing them to fill a very different niche. They developed longer legs, shorter toes, longer Achilles tendons, wider shoulders and a stronger gluteus maximus for running[1]. In addition, they evolved larger weight-bearing joints that could support long runs while avoiding too much damage.

It is likely that humans also have a much higher metabolic capacity than our ancestors did, measured by our ability to take up and use oxygen (VO_2max). We are the only primates with the aerobic capacity to support long-distance running. It is otherwise restricted to migratory ungulates (horses, wildebeest) and social carnivores (hyenas, wolves). Humans evolved from animals with a low VO_2max to the modern endurance athlete's maximum capacity approaching 90 millilitres of oxygen per kilogram of body mass per minute.

Exercise remains central to our basal metabolism, even though it is no longer a core part of life for some human populations. For example, persistence hunting led to humans having great capacity to conserve sodium and retain water when either are in short supply[3].

We think that one protein could have been central to these dramatic physiological changes, both mental and physical: brain-derived neurotrophic factor (BDNF), which increases with exercise and play, and sculpts the developing brain (see 'Memory jog').

Although some research has yielded mixed results about the benefits of exercise for the brain[5], the literature in general shows that regular exercise can have many cognitive benefits. Some research suggests that exercise has antidepressant effects, at least against mild depression, and may even offset Alzheimer's disease: an exercise programme has been shown to increase the volume of the brain's hippocampus. More data are needed to obtain a clearer picture of the effects of exercise on cognition. Ongoing clinical trials are testing the effects of exercise on Alzheimer's disease and schizophrenia.

EXERCISE FOR LIFE

We believe that at least some of these beneficial effects are mediated by BDNF. First, circulating BDNF is increased by exercise, partly by release from the nerve–muscle junction into the blood. In muscle, BDNF can increase protein synthesis and fat metabolism[5,6], which are key targets of exercise. Mice that lack BDNF become obese and there are strong inverse links between BDNF and type 2 diabetes[7] — a disease for which it is well known that exercise, with the appropriate dietary changes, is the best therapy allied to medication.

In the brain, BDNF increases neuronal connections and is crucial for some aspects of memory[6]. BDNF plays a part in the hypothalamus, controlling body weight and energy homeostasis[7]. It also triggers the brain mitochondria — the powerhouses of the cell — to use oxygen more efficiently[8], in a similar way to how exercise helps to increase overall VO_2max in humans.

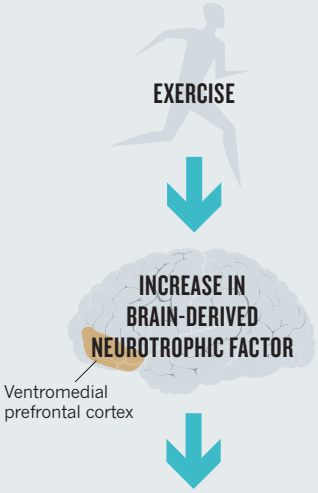
As humans needed more brain power to track prey, increases in BDNF may have helped to build up the hippocampus and prefrontal cortex — key brain areas that are involved in spatial mapping, decision-making and control of context, fear and emotions, including violence[8–10]. BDNF has a crucial role in the prefrontal cortex, a region that is also strongly associated with psychiatric disorders[9]. Not surprisingly, BDNF is reduced in hippocampal and cortical regions in models of stress and psychiatric disorders[10], as well as in Alzheimer's disease. Putting it all together, we think that exercise increases BDNF in key areas of the brain, which, in turn, has physiological effects that help to protect humans from psychiatric problems, including violent behaviour.

Not all of the health benefits of exercise stem from the effects of BDNF alone, because diabetes and obesity are known to have other, multiple causes. Importantly, the modern diet — which is far removed from the whole foods our ancestors ate — can itself lead to obesity and inflammation, which counteract many of exercise's benefits.

A lack of exercise may have multiple, long-term damaging effects, particularly when coupled with a poor diet. Simple exercise programmes in schools and sports clubs, for example, are probably the most cost-effective investments a society can make in its psychological and physical health. Exercise may be cheap, but the consequences of ignoring it are costly. ■

MEMORY JOG

The beneficial effects of exercise on the body and brain are increasingly thought to be mediated by the brain protein BDNF.



- Improves metabolism.
- Opposes neurodegenerative processes.
- Acts on spinal cord to reduce heart rate.
- Reduces rates of obesity and type 2 diabetes in mice.
- Increases connectivity between neurons.
- Triggers brain mitochondria to use oxygen more efficiently.
- Has a role in psychiatric and neurological disorders.

Timothy Noakes *is professor of exercise and sports science at the University of Cape Town, Rondebosch 7701, South Africa.*
Michael Spedding *is a pharmacologist and lives at Le Vésinet, near Paris, France.*
e-mails: timothy.noakes@uct.ac.za; spedves2@orange.fr

1. Bramble, D. M. & Lieberman, D. E. *Nature* **432,** 345–352 (2004).
2. Konopka, G. & Geschwind, D. H. *Neuron* **68,** 231–244 (2010).
3. Noakes, T. *Waterlogged: The Serious Problem of Overhydration in Endurance Sports* (Human Kinetics, 2012).
4. Hill, K. R. *et al. Science* **331,** 1286–1289 (2011).
5. Chalder, M. *et al. Br. Med. J.* **344,** e2758 (2012).
6. Mattson, M. P. *Ageing Res. Rev.* **11,** 347–352 (2012).
7. Noble, E. E., Billington, C. J., Kotz, C. M. & Wang, C. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **300,** R1053–R1069 (2011).
8. Markham, A. *et al. Eur. J. Neurosci.* **35,** 366–374 (2012).
9. Agid, Y. *et al. Nature Rev. Drug Discov.* **6,** 189–201 (2007).
10. Qi, H. *et al. Neuropharmacology* **56,** 37–46 (2009).

M.S. declares competing financial interests; see go.nature.com/rhkkl5.

Genetically enhanced Olympics are coming

Future Olympic Games may allow handicaps and gene therapy for people born without genes linked to athleticism, predict **Juan Enriquez** and **Steve Gullans**.

Olympians can run faster, leap higher and lift more than 'normal' humans. Of course, such elite athletes earn their titles with an astonishing amount of hard work and support. But many also have some unearned advantages: the right genes.

There is growing evidence that world-class athletes carry a minimum set of particular 'performance-enhancing' genes. For instance, almost every male Olympic sprinter and power athlete ever tested carries the 577R allele, a variant of the gene *ACTN3*. About half of Eurasians and 85% of Africans carry at least one copy[1] of this 'power gene'. The billion or so other people who lack the 577R allele might wish to reconsider their Olympic aspirations.

More and more genes are now being linked to athletic prowess, and future Olympic officials will have to wrestle with the implications. Are the games in fact a showcase for hardworking 'mutants'? And if Olympic rule-makers admit that the genetic landscape is uneven, should they then test every athlete and hold separate competitions for the genetically ungifted?

There are three basic scenarios for future Olympics. First, the competition could continue as a showcase of athletes born with genetic advantages. Another option would be to use handicaps — similar to those that are now used in several non-Olympic sports — to level the playing field for athletes who do not carry beneficial genes. A third option, if safe, would be to allow athletes who did not win the genetic lottery to 'upgrade' through gene therapy — a practice that is now banned as 'gene doping'.

We have been living in the first scenario for centuries. More than 200 gene variants are already associated with athleticism[2]. For example, carriers of the 'I' variant of the gene *ACE* are more likely than non-carriers to successfully climb an 8,000-metre peak[3]. The I variant is present in 94% of Sherpas in the Kathmandu Valley of Nepal[4], but in only 45–70% of people of other ethnicities[5]. It is associated with increased endurance. A study of British runners found that it is most common in those who race the longest distances[6].

Such variants occur frequently in the human population, and athletes probably need a subset of them to achieve elite status. As more individual genomes are sequenced, researchers will begin to detect some rare variants that differentiate truly superior champions from mere world-class athletes. Eero Mäntyranta had a mutation in the gene *EPOR* that caused him to produce extra red blood cells, boosting his oxygen-carrying capacity by 25–50% (ref. 7), which probably helped him to earn seven Olympic cross-country ski medals.

But how easily could scientists detect whether a variant is natural or introduced? Even 'gender-verification' testing to confirm the sex of female competitors has been problematic, given the natural biological variation among individuals[8].

ILLUSTRATION BY GARY NEILL

LET THE GAMES BEGIN
Science and the Olympics
nature.com/olympics2012

Olympic traditions change glacially, but eventually, what was once unthinkable becomes commonplace. Once upon a time, women were allowed to compete only in Olympic tennis, golf and croquet. Until the 1970s, paid athletes were banned from Olympic competition — now, professional basketball players compete for medals. And 'extreme sports' such as snowboarding and bicycle motocross have now become Olympics-worthy.

As officials struggle with the implications of genetic data and upgrades, we will probably see, initially, a set of draconian rules against gene modification. Will a competitor who was cured of sickle-cell anaemia by gene therapy as a child be excluded? How about someone cured of an *EPOR* defect through use of Eero Mäntyranta's natural variant?

Just as Oscar Pistorius, the Paralympic champion runner who was once banned from the Olympics because he uses leg prostheses, will now compete in London on the South African relay team, we expect that as genetic modification becomes more common, a gradual acceptance of safe genetic enhancements will follow. After all, we watch the games today to marvel at athletes who are 'faster, higher, stronger' — whether man or woman, amateur or professional, 'disabled' or not. ■

Juan Enriquez *and* **Steve Gullans** *are managing directors of Excel Venture Management, Boston, Massachusetts 02199, USA, and the authors of* Homo Evolutis: Please Meet the Next Human Species *(TED Conferences, 2011).*
e-mail: jenriquez@excelvm.com

1. Berman, Y. & North, K. N. *Physiology* **25,** 250–259 (2010).
2. Ostrander, E. A., Huson, H. J. & Ostrander, G. K. *Annu. Rev. Genomics Hum. Genet.* **10,** 407–429 (2009).
3. Thompson, J. *et al. High Alt. Med. Biol.* **8,** 278–285 (2007).
4. Droma, Y. *et al. Wilderness Environ. Med.* **19,** 22–29 (2008).
5. Sonna, L. A. *et al. J. Appl. Physiol.* **91,** 1355–1363 (2001).
6. Myerson, S. *et al. J. Appl. Physiol.* **87,** 1313–1316 (1999).
7. de la Chapelle, A., Träskelin, A. L. & Juvonen, E. *Proc. Natl Acad. Sci. USA* **90,** 4495–4499 (1993).
8. Karkazis, K., Jordan-Young, R., Davis, G. & Camporesi, S. *Am. J. Bioeth.* **12,** 3–16 (2012).

W. LOVELACE/GETTY

British gangster Ronnie Kray (right) was bisexual, unlike his twin, Reggie (left). The disparity could be down to epigenetics.

EPIGENETICS

Different under the skin

Michael G. Sargent enjoys a discussion of twin studies that aims to unpick the effects of nurture on nature.

Human personality traits apparently governed by genes may change in response to crucial life experiences. But how? Through epigenetics — chemical modifications of the genome and certain associated proteins. In *Identically Different*, genetic epidemiologist Tim Spector argues that identical twins offer a unique opportunity to understand this mysterious process.

Far from presenting a put-down of a gene-centred view of heredity, Spector believes that the influence of genes is paramount, but that identical-twin studies reveal possibilities for variation. He focuses on twin studies because they are a well-trodden way of exploring how environmental triggers can initiate a chronic disease such as rheumatism in one twin by an epigenetic process, while the DNA sequence remains constant.

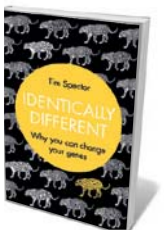
Spector also seeks to identify environmental cues that affect personality traits — criminality, talent, homosexuality, fidelity, autism and many others — that have sometimes been simplistically ascribed to specific genes. (With some irony, Spector names his chapters after phrases used by the media — such as 'the fat gene' or 'the gay gene'.) Although the differences between twins cannot yet be attributed to specific epigenetic processes, animal-behaviour studies suggest that this is plausible.

NATURE.COM
For another review on epigenetics, see:
go.nature.com/ngbo5g

A number of the case studies in *Identically Different* are of identical twins who were separated at birth, but grew up with spookily similar behaviour. The main focus, however, is on separated twins who showed important differences, such as Nina and Gill.

Nina grew up an only child in privileged circumstances. Gill had five siblings in a rough-and-tumble home with no encouragement or material benefits. As teenagers, both were wayward, fell pregnant and were married very early to abusive men. Nina quickly divorced her tormentor, went to university and lived happily thereafter; Gill endured many years of misery before she escaped, completed her education, became a civil servant and married happily, albeit dogged by significant health problems that were perhaps related to neglect. So the huge difference in upbringing did not affect their personality, naughtiness as children or susceptibility to temptation — but decisively affected their education, confidence and finances.

Spector observes that the influence of parental background rarely outweighs the genetic legacy. Occasionally, however, particular individuals make a crucial difference. Often these are teachers: small differences in encouragement perceived by each twin can markedly affect educational progress.



Identically Different: Why You Can Change Your Genes
TIM SPECTOR
Weidenfeld & Nicolson: 2012. 336 pp. £20/$25.10

Twin studies show that most kinds of talent have a genetic element. But at the highest levels, only one twin tends to emerge as a star — as with actor Isabella Rossellini or car-racing champion Mario Andretti. Curiously, twins seem to avoid competing directly, which is usually attributable to a difference in motivation and hard work, perhaps influenced by lucky or unlucky random events.

Spector reports that a strong genetic component to criminality emerges from investigations of identical twins, but a history of abuse

is also a factor in children drifting towards a life of crime. In one study, those carrying a specific gene variant who were also victims of abuse were nine times as likely to become delinquents than carriers who were not abused. The variant encoded a defective version of monoamine oxidase, an enzyme that regulates levels of certain neurotransmitters.

The significance of genetic factors in criminality is evident in Spector's observation that only one-third of identical-twin pairs who experienced childhood abuse are both inclined towards criminality. Magnetic resonance imaging of the brains of individuals in that criminal subgroup revealed an excess of grey matter — generally regarded as a sign of brain immaturity, and also seen in psychopaths and individuals with autism spectrum disorder. Spector suspects that this pathology involves the epigenetic modification of genes that profoundly affect behaviour, such as the stress response, mood-regulating neurotransmitters and the "trust hormone" oxytocin.

Spector believes that sexual preferences are governed by a substantial genetic factor, but there are many instances in which one twin is straight and the other gay — notably Ronnie Kray, the British gangster, who was bisexual, unlike his brother Reggie, and also had schizophrenia. One possibility is that sexual orientation relates to prenatal hormonal exposure, which affects brain development. It also leaves a curious anatomical signature: in heterosexual males, the index finger tends to be shorter than the ring finger. The reverse is true in gay men.

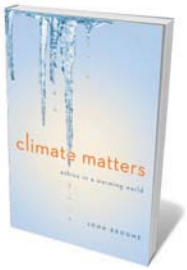
Perhaps the strangest idea to emerge from twin studies is that belief in a deity may have a substantial genetic component. This seems to transcend particular faiths and, mysteriously, maps to a region of chromosome 15 that lacks any protein-coding sequence.

Spector skates over the biochemistry of epigenetics, without reference to recently recognized players, such as microRNAs, that might modify neuronal activity. More discussion about different sorts of identical twin might have been informative. Are twins who shared a placenta more similar than those who did not? Are 'mirror-image' twins — those with small asymmetries in appearance — more different than those who are truly identical?

Real case histories of identical twins may be the only way to help us to understand how life experiences influence personality and behaviour. They may also suggest how some problems might be more manageable than we had imagined. ■

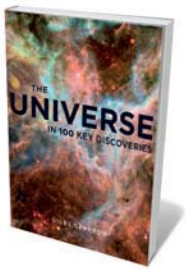
Michael Sargent *is a developmental biologist at the National Institute for Medical Research, Mill Hill, London, and author of* Biomedicine and the Human Condition: Challenges, Risks and Rewards. *e-mail: msargen@nimr.mrc.ac.uk*

Books in brief

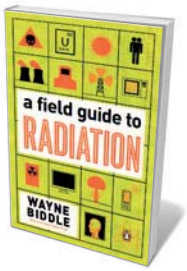


Climate Matters: Ethics in a Warming World
John Broome NORTON *224 pp. £14.99 (2012)*
With climate-change policy looking increasingly toothless, we need fresh ways of grappling with this environmental catastrophe. Philosopher and "lapsed economist" John Broome vaults in where policy-makers fear to tread, exploring the moral aspects of climate choices. In the latest instalment in the Amnesty International Global Ethics Series, Broome argues that countries and individuals are ethically obliged to curb emissions. With penetrating clarity, he uses science and economics as a springboard to cover big issues, from the need for action despite uncertainty to the value of human life.

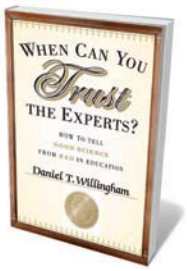
Beyond the Blue Horizon: How the Earliest Mariners Unlocked the Secrets of the Oceans
Brian Fagan BLOOMSBURY *336 pp. £20 (2012)*
What motivated the first seafarers to take to uncharted open water — land grabs, a thirst for trade, conflicts at home or restlessness? Anthropologist Brian Fagan sails far back, beyond explorers such as Magellan and Cook, to when those intrepid pioneers travelled in rafts, coracles and longboats. Starting 50,000 years ago with the southeast Asian exodus to the Pacific Islands, he also examines early sailors in the Aegean Sea, monsoon winds, Norse voyages and the complexities of marine exploration in the ancient Americas.



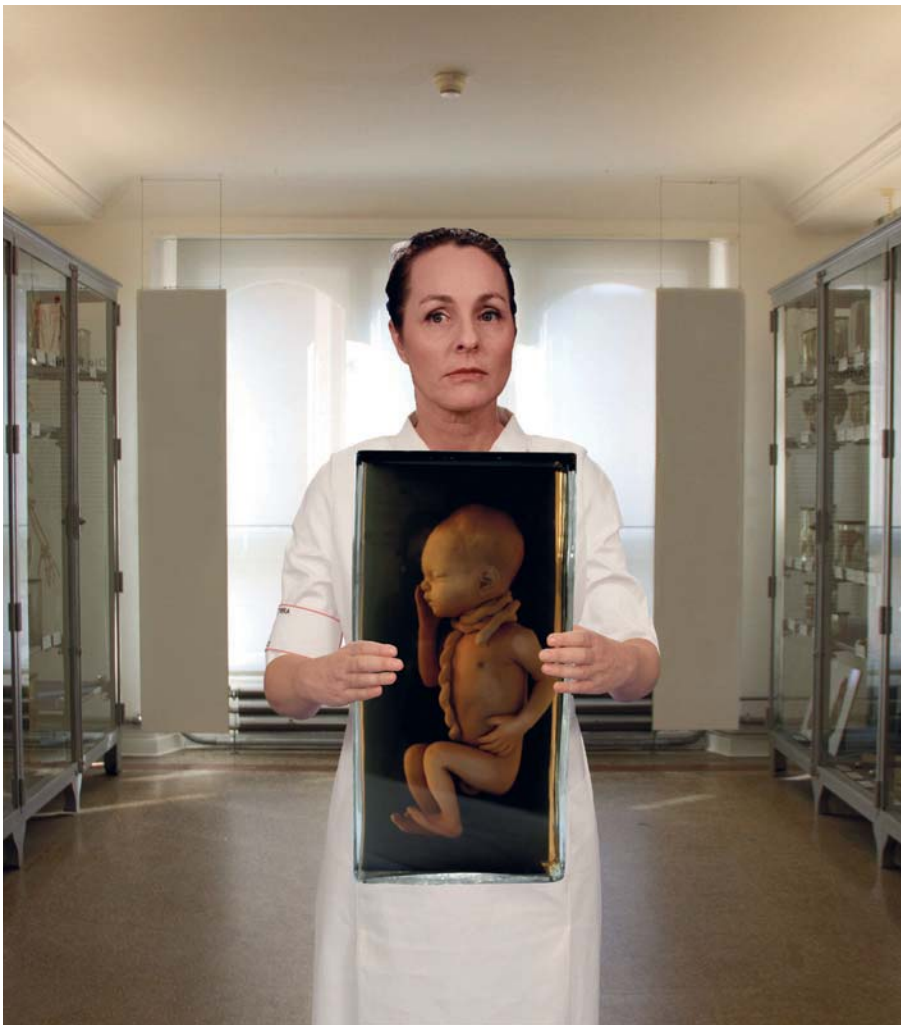
The Universe: In 100 Key Discoveries
Giles Sparrow QUERCUS *416 pp. £19.99 (2012)*
If you hanker for a compact compendium of cosmological breakthroughs, this is it. Astronomy writer Giles Sparrow is an able guide through 100 discoveries that have shaped understanding of the Universe and its workings. Trawling the eons, we explore the scientific revolution that unseated Earth from the centre of the Universe, the bombardment of Earth 4 billion years ago, flare stars and much more, finishing with speculation about the cosmological endgame. Essays and stunning images are framed by a definition and description of each breakthrough and its relative importance.



A Field Guide to Radiation
Wayne Biddle PENGUIN *288 pp. £10.23 (2012)*
Pulitzer prizewinning writer Wayne Biddle, author of the award-winning *A Field Guide to Germs* (Henry Holt, 1995), here tackles another ubiquitous aspect of daily life: radiation. He briefly covers the history — pioneering researcher Marie Curie, to whom the "glowing tubes looked like fairy lights", the stockpiling of nuclear warheads and the spread of nuclear power — before moving on to radioactive elements and related phenomena, from critical mass and decay products to fallout and occupational radiation. Witty, succinct and handily organized in an A–Z format.



When Can You Trust the Experts?: How to Tell Good Science from Bad in Education
Daniel T. Willingham JOSSEY BASS *272 pp. £16.99 (2012)*
Cognitive psychologist Daniel Willingham offers a cautionary tale about poor science in education. With some teaching tools backed by research that is far from robust, Willingham calls for a four-step process for selecting the best of them: 'strip it' (look at the claim and decipher the promised outcome); 'trace it' (find the source of the idea and how others view it); 'analyse it' (determine whether the evidence is sound); and ask, 'should I do it?' (factor in the urgency of the need).



AANDO FINE ART, BERLIN/A. BEN RON

In *The Cord*, Aya Ben Ron reimagines the history of a baby accidentally strangled during birth.

PATHOLOGY

The dead reborn

Artworks confronting the ethical dimensions of pathology specimens intrigue **Alison Abbott**.

If you are drawn to disembodied human organs floating in formalin, the Berlin Museum of Medical History at the Charité university hospital is just the place to visit. But behind the cases is a multifaceted political and scientific history. Into that steps Israeli artist Aya Ben Ron, whose show *A Voyage to Cythera* has created an 'intervention' — a quiet imposition of her own way of seeing — among the pathology specimens.

There are some 750 of these on permanent display — a fraction of the tens of thousands put together by pioneering pathologist Rudolf Virchow (1821–1902) and his successors. Virchow called the collection his 'dear child' and died long before it was decimated by Second World War bombs. After the war, the museum found itself in communist East Germany, next to the Berlin Wall, with its upper west-facing windows painted black. The reunification in 1990 then saw specimens reorganized according to modern aesthetic tastes and scientific standards, aiming for dignity and didactic power.

Ben Ron's artworks — sculpture, video, sound and poetry — 'intervene' at 18 stations in a viewing trail (http://avoyagetocythera.com). Appearing dressed as a nurse in the videos, the artist presents the specimens from the viewpoint of a carer. For her, they cannot be Virchow's dear child. Nor, without the dimension of emotion, can they be truthful teaching aids: each was once part of a person. The names of these people may be long forgotten, but Ben Ron has invented new histories for them.

A Voyage to Cythera
The Berlin Museum of Medical History. Until 9 September 2012.

The title of the exhibition is taken from the 1857 poem by Charles Baudelaire, in which the protagonist, in search of the goddess Venus, travels to the Greek island of Cythera, her birthplace. Instead of finding the goddess of love, he discovers, hanging from a tree, a rotting carcass — which turns out to be his own.

At its most basic level, this thoughtful body of works reflects, like Baudelaire's poem, on the natural yearning for love and beauty, and the reality of life — which, at the end of the day, can offer only the opposite. Ben Ron laser-cuts her sculptures (of bandaged figures or anatomy classes) from mild steel, so they begin to rust slowly from the moment of their creation.

Two of the videos focus on specimens that Ben Ron selected for their particular pathos; she invents their history in poems. One is a vast distended colon from a person who died, essentially, of constipation. Ben Ron dignifies the illness through a metaphorically rich poem. As nurse, she cradles the monstrosity protectively, and with tenderness.

The other is a perfectly developed baby, tragically strangled by its umbilical cord during birth. Both the images and poem are heart-wrenching, yet also somehow heart-warming: the child is, Ben Ron writes, "if not alive, at least in a jar".

In a room housing cardiovascular specimens, she presents a recording of a heartbeat. It sounds infrequently: a low, doom-ridden boom providing a visceral reminder to visitors that the hearts they gawp at were once alive. The sound of summer meadows plays low in the background of a room containing other organs. The birdsong lifts the spirits until the sudden, brutal buzz of a bluebottle — which lays its eggs on decaying flesh — flips the mood.

One station is adjacent to a west-facing window. No longer blacked out, its clear glass looks out onto Berlin's new central station with its networks of tracks reaching across Europe. Germany's reach was not always benign; Ben Ron was influenced by her grandmother, a gynaecologist and Holocaust survivor. One of the museum's rooms is dedicated to detailing the abuses of medicine under the Third Reich. Ben Ron has a station there: a poster that hijacks the name of a 1941 propaganda film legitimizing the murder of the disabled or incurable: 'I accuse.' ■

Alison Abbott *is Nature's Senior European Correspondent.*

SYFY

In *Alphas*, Ryan Cartwright plays Gary Bell, a superhero with autism and the ability to 'read' electromagnetic signals.

Q&A Bruce Miller

Superpower sleuth

The US television series Alphas *features an unusual breed of superhero: ordinary people with extreme abilities. In the run-up to the second season, head writer Bruce Miller explains how he sifts through the latest scientific findings to craft an array of superpowers.*

ILLUSTRATION: NICK HIGGINS
PHOTO: P. ECCLESINE/WARNER BROS/GETTY

Where does *Alphas*' particular mix of fact and the fantastical come from?
Zak Penn and Michael Karnow created *Alphas*; they have a lot of experience in the world of comic books. They were looking for that sense of realism, of being grounded in our world. Our show is much more fulfilment than fantasy. I relate it to perfect pitch. I have no ear for music, and my son has perfect pitch; it might as well be a superpower as far as I am concerned. For the show, we try to extrapolate in one direction or another from our knowledge of neurology and from that of authors such as Oliver Sacks, who study brains that work in a slightly different way. For example, if you could live for 200 years and could heal yourself, what would happen to your brain — are there parts that couldn't keep up?

How do you develop a new character?
When we started work on the second season, we were fascinated by short- and long-term memory, how long-term memories are stored differently and where they are stored. We asked, "What is everybody shooting for in their lives?" So many people are trying to live in the moment, so we came up with a character who embodies the good and the bad of that: Kat, played by Erin Way. We had read about muscle memory, for instance: how, after repeated movement, the transmission of nerve signals to the muscles involved becomes more effective, and those movements become automatic. Kat has instant muscle memory — and can also learn anything just by reading a book on the subject. But things that happened to her more than a month ago start to fade away.

What kind of people make up your writing staff?
It is a mix. We have people who know the title, the running time and the original broadcast date of every episode of *Star Trek*. But my affection for this show and genre came much more from science. I think I live somewhere in the *Wired* world — a little more science than science fiction.

Do you have science advisers?
We have a wonderful science adviser at the University of California, Los Angeles: the neurologist Susan Bookheimer. She gives us a lot of guidance on autism, because we have an autistic character, Gary, and we try to make him as real as possible. She reads every script and tells us, say, when Gary wouldn't use a metaphor in a certain way. There is an episode later in the second season in which Gary has to take care of a foundling child, for instance. Susan spoke of the desire someone like Gary would have to bond with a child, and how Gary might try to satisfy the needs of a baby — research he would do, and so on. We also spoke about the current studies using drugs, such as vasopressin, to enhance social affiliation in autistic children.

Can you discuss the scientific phenomena from the forthcoming season?
We have an episode about infrasound — very low-frequency sound that causes hallucinations. We also have a character who moves much more quickly than anyone else. He is all about perception, plus fast-twitch muscles plus circadian rhythm — that is, what if someone had a different biological clock to the rest of us? He is 22 but he looks like he is in his forties. This works more as metaphor than science: it could be the stress of his condition that makes his body age, or he might actually be going through life faster. We tend to think that the way we perceive time is absolutely the way time goes by, but there can be vast differences in individual experiences of time's passing, and also between people's experiences and our perceptions of them.

Does science ever outpace the programme?
Sometimes we read about something and start to work on a script. Then, three months later, when we've finished the script, we will find it is already happening — we learn about it as scientific papers are picked up rapidly online and in the popular press. And so we say: "Oh, it's not fascinating enough — we need to push it a little."

INTERVIEW BY MARC WEIDENBAUM

Correspondence

Open access: let's go for gold

The report from the independent Finch group in the United Kingdom recommends a 'gold' route to open-access papers, in which journals impose pre-publication charges, over a 'green' route, in which manuscripts are made available through a repository (see *Nature* **486,** 439; 2012). As secretary to the Finch committee that authored the report, I can explain why we believe that gold is the way forward as the main vehicle for publishing research.

Open access by the green route is permitted only after an embargo period, and only to an unpublished version of the paper, without links or semantic enrichment for web applications. Rights of use and re-use are severely restricted. Commercial and not-for-profit journals that rely on subscriptions impose these constraints to protect their income. Green open access without any restrictions cannot work alongside subscription-based publishing.

Gold open access avoids these problems: journals receive their revenues up front, so they can provide immediate access free of charge to the peer-reviewed, semantically enriched published article, with minimal restrictions on use and re-use. For authors, gold means that decisions on how and where to publish involve balancing cost and quality of service. That is how most markets operate, and ensures that competition on quality and price works effectively. It is also preferable to the current, non-transparent market for scholarly journals.

The main barrier to gold open access has been a lack of systematic payment arrangements for article-publishing charges. The UK research councils should follow the Wellcome Trust in providing straightforward and flexible payment mechanisms. The costs would represent a small rounding error set against current levels of expenditure on research.

Green open access may be cheaper, but that is not the point: watching your favourite football team playing live beats seeing the highlights later on television.

Michael Jubb *Research Information Network, London, UK.*
michael.jubb@researchinfonet.org

Open access: a green light for archiving

The Finch report (see *Nature* **486,** 439; 2012) on ways to achieve universal open access to scientific papers favours a 'gold' approach that involves a publication charge for authors. This must not stand in the way of 'green' self-archiving of open-access journal articles (see go.nature.com/gvjwiw).

Research institutions and funders need to mandate green open-access self-archiving of the final, refereed versions of all journal articles as soon as they are accepted for publication.

Paying for gold open access pre-emptively today, without funders or institutes having first mandated green open access, will delay universal access and waste scarce resources. The money to pay for gold open-access publishing will become available once green open access has eventually made subscriptions unsustainable.

Stevan Harnad *University of Quebec at Montreal, Canada, and University of Southampton, UK.*
harnad@ecs.soton.ac.uk

Data sharing is harder to reward

Open science has won another powerful advocate in the UK Royal Society (*Nature* **486,** 441; 2012). But freely sharing research results can have social repercussions that may be damaging to science.

By confusing the allocation of scientific merit and potentially undermining authorship conventions (see, for example, T. Rohlfing and J.-B. Poline *NeuroImage* **59,** 4189–4195; 2012), data sharing could work against individual scientists' need for recognition. This is one reason why scientific institutions, from universities to research councils, do not reward data sharing.

Policy-makers need to remember that, right or wrong, competition has always been a strong driver of science.

Gerrit Hirschfeld *German Paediatric Pain Centre, Children's and Adolescents' Hospital, Datteln, Germany. g.hirschfeld@kinderklinik-datteln.de*

Costa Rica pioneers ecosystem services

As the world waits to see how the United States and China respond to the prevailing environmental, social and economic crises, the small country of Costa Rica can offer some sizeable lessons.

Costa Rica has been pioneering a green economy for 15 years through its Payments for Environmental Services (PES) programme. Under this scheme, landowners who maintain environmental services are rewarded by the people who benefit from them. For example, payments are levied from water users to pay highland landowners who keep the water flowing downstream by planting and protecting forests.

The country's 1996 Forest Law strengthened regulations for existing forests, defined ecosystem services, clarified property rights to allow trading of environmental services, and created the National Forestry Financing Fund (FONAFIFO) to oversee PES payments. Funds are drawn from taxes on fossil fuels and on water.

Voluntary payments by individuals to a national bank offset vehicle emissions. People can also use a 'green' debit card that transfers 10% of the bank's commission to FONAFIFO.

Costa Rica's commitment to becoming carbon-neutral by 2021 is encouraging local voluntary agreements and parallel schemes such as PSA Solidario (go.nature.com/4lto1s), which rewards small-scale farmers who conserve forests but cannot access the national programme.

Ina Porras *International Institute for Environment and Development, London, UK.*
ina.porras@iied.org

Sewage recycles antibiotic resistance

Your discussions on toilet technology (*Nature* **486,** 185; and *Nature* **486,** 186–189; 2012) should have mentioned an important aspect of sewage disposal — the problem of contamination with antibiotics and antibiotic-resistant organisms.

Processes for treating waste water provide perfect mixing pots for bacteria carrying genes that confer antibiotic resistance. These genes can come from many environmental sources and are often conveyed on mobile genetic elements. Ways must be urgently sought to prevent the genetic recycling and dispersal of antibiotic-resistance genes and of resistant organisms.

Julian Davies *University of British Columbia, Vancouver, Canada.*
jed@mail.ubc.ca

CORRECTION

The Obituary for David Sayre (*Nature* **484,** 38; 2012) incorrectly implied that he suggested focusing X-rays using Fresnel zone plates. In fact, his contribution was the use of nanofabrication to make the plates.

GRADUATE STUDENTS Union membership could reopen for US research assistants **p.397**

EARLY-CAREER RESEARCHERS Global coalition to advocate for better conditions **p.397**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

B. WELSH/CORBIS

RISKY RESEARCH

The sky's the limit

Transformative research projects can bring big rewards. But securing funding requires a particular set of strategies.

BY VIRGINIA GEWIN

Miguel Nicolelis has made advances that could help people with paralysis to walk again. That success was possible thanks to funding earmarked for high-risk, high-reward research. "Usually you have to write a grant on a narrow project using a technique you are deemed an expert in, but that's not how major discoveries occur — for that, you have to explore a vision," says Nicolelis, a neurobiologist at Duke University in Durham, North Carolina. His vision has him recording the activity of large populations of neurons and developing a theory of how brain circuits work. With that, he translates the brain's electrical activity into digital signals that a robotic suit can interpret to control body movements.

These bold breakthroughs grabbed public attention — and earned Nicolelis an appearance on popular US television programme *The Daily Show* in March last year. "This is our moonshot," Nicolelis told host Jon Stewart.

But he says that a proposal to enable a monkey to control an on-screen avatar using brain electrodes would have been laughed out of the room by the review panel for an R01, the standard US National Institutes of Health (NIH) grant for biomedical research. Luckily, says Nicolelis, reviewers for the NIH Director's Pioneer Awards, a funding scheme to support bold scientific leaps rather than incremental advances, saw the potential of his work. His research not only earned him a Pioneer award, but also led to new research exploring a treatment for Parkinson's disease using minimally invasive spinal-column stimulation with an NIH-funded Transformative R01 award.

Many scientists are concerned that the conventional grant-review system has become too conservative, and that this trend has been exacerbated by budget crunches in recent years. All too aware of this perception, funders are creating new types of grant schemes (see 'In support of innovation'). Many such schemes aim to bring together interdisciplinary brainstorming teams to tackle the world's big problems.

Innovative grant mechanisms put less emphasis on primary data than on vision, imagination, reasoned logic and relevance to global issues. Not all researchers are equipped for such a shift in strategy, but those eager to break new ground would be wise to adjust their thinking.

ON FURTHER REVIEW

Funding schemes identify the most promising risky research in different ways. Four programmes run by the European Research Council, collectively funded at €1.75 billion (US$2.15 billion) for 2013, rely on conventional peer review, but do not pre-select topics; investigators can identify ideas free of political, geographical or economic considerations.

Other schemes use more unorthodox methods. Last year, the US National Science Foundation (NSF) announced the US$24-million Creative Research Awards for Transformative Interdisciplinary Ventures (CREATIV) pilot programme, which awards grants through an internal process without peer review. Researchers don't simply submit a proposal: they must first send an inquiry, authorized by directors from two intellectually distinct NSF programmes. "It's a pretty small fraction of inquiries that lead to a proposal," says Tom ▶

PIONEERING PROGRAMMES

In support of innovation

Governmental funding bodies are recognizing the need for mechanisms that encourage blue-skies research. Here are some of the most popular.

EUROPEAN RESEARCH COUNCIL (ERC)
Starting Grants — for researchers who have completed their PhDs in the past 2–7 years.
Consolidator Grants — for researchers who have completed their PhDs in the past 7–12 years.
Advanced Grants — for researchers at any career stage doing frontier research.
Synergy Grants — for interdisciplinary teams of 2–4 principal investigators.
Proof of Concept — for grant-holders developing innovations from their ERC-funded frontier research.

US NATIONAL INSTITUTES OF HEALTH
Early Independence Award — for new PhD holders eager to skip postdoc training and start an independent laboratory.
New Innovator Award — for creative early-career investigators who lack the preliminary data necessary for conventional grants.
Pioneer Award — for scientists at any career stage, who will spend at least 51% of their research effort on the pioneer research.
Transformative R01 — for bold, paradigm-shifting but untested ideas from teams or individual researchers at any career stage.

US NATIONAL SCIENCE FOUNDATION (NSF)
Creative Research Awards for Transformative Interdisciplinary Ventures (CREATIV) — proposals must include approval from two intellectually distinct NSF divisions or programmes.
Early Concept Grants for Exploratory Research (EAGER) — for untested but potentially transformative ideas.
Special creativity extensions — for high-risk opportunities not covered by the proposal for a standard grant. Based on the recommendation of the relevant programme officer.

US ADVANCED RESEARCH PROJECTS AGENCY—ENERGY
Open solicitations for transformational technologies — for early-stage energy-research projects from any discipline that would not attract private investment.
Targeted solicitations — for research projects on agency-determined topics ranging from high-energy advanced thermal storage to materials for advanced carbon-capture technologies. V.G.

▶ Russell, programme director for CREATIV in Arlington, Virginia. "The vast majority of inquiries are not promising or appropriate, and the programme directors act as a tough filter."

Applications to the NIH's high-risk, high-reward programme — encompassing Pioneer awards and Transformative R01s, among other schemes — have a 5% success rate, and go through a two-panel review process. Pioneer proposals first go to three generalist reviewers who have a broad view of science and are not allowed to discuss the applications with each other. A second panel scores the resultant reviews and selects the 25 most exciting projects. Proposals don't require data or a detailed research plan; applicants need only suggest how they will accomplish the research, and describe their qualifications and how they have overcome research roadblocks. "These aren't incremental awards; these are big ideas that move the field forward," says James Anderson, director of programme coordination, planning and strategic initiatives at the NIH in Bethesda, Maryland.

Without the need to include preliminary data, applicants must think differently about what they write. "It's hard to make yourself write those kinds of proposals," says Eric Toone, principal deputy director at the US Advanced Research Projects Agency—Energy (ARPA-E) in Washington DC, launched by the government in 2010 to encourage risky, transformative ideas. Funders recognize that speculative research may not pan out, but they want to see radical ideas that will yield interesting insights. Researchers are often so entrenched in incremental approaches — or hindered by the need to secure tenure — that, say programme officers, it takes time for them to work out how best to write radical proposals. "I suggest that researchers try reframing problems," says Tina Seelig, executive director of the Stanford Technology Ventures Program at Stanford University in California, and author of *inGenius: A Crash Course on Creativity* (HarperOne, 2012). In biomedical research, for example, they could "reframe the questions in terms of wellness rather than sickness".

It is that original perspective that funders are looking for. "We want to have that 'holy cow' moment when reading a proposal, one that makes clear the potential to change how we think about a technology area," says Toone. A researcher's enthusiasm for a high-risk project can make or break the case for funding. "A good grant reads like a novel; it grabs you on the first page and you can't put it down," says Nicolelis.

Applicants also need to convey the potential

impact of the project, says Ravi Basavappa, NIH programme manager for high-risk, high-reward funding. "Why is this proposed project so important, what communities would be affected and how?" he asks.

Funders suggest that researchers discuss their ideas with the appropriate programme directors before submitting a grant. "Find the people running those programmes and see if you catch their interest with a description of what you want to do," advises Russell.

NEW APPROACHES

Some funders are going even further off the beaten path. At the NSF, programme directors can authorize a 'special creativity extension' to fund work not covered under a standard grant. In 2010, the UK Biotechnology and Biological Sciences Research Council (BBSRC) in Swindon teamed up with the NSF to create a jointly funded Ideas Lab: a five-day meeting to brainstorm ways to improve plant photosynthesis and enhance food production. Participants hashed out the most promising approaches and wrote proposals that were reviewed at the meeting; the funding agencies shared a total of £6.15 million (US$9.5 million) between the best projects. Another Ideas Lab is planned for later this year, this one on producing crops that require less nitrogen fertilizer.

The UK Engineering and Physical Sciences Research Council in Swindon also supports high-risk work. For its Bright IDEAS Awards, it offers researchers up to £250,000 over 18 months to tackle a specific challenge — most recently, the development of quantum technologies that could transform communication, imaging or computing.

DUKE UNIV./J. WALLACE

"A good grant reads like a novel; it grabs you on the first page and you can't put it down."
Miguel Nicolelis

The Research Corporation for Science Advancement (RCSA) in Tucson, Arizona, runs the Scialog programme, in which it provides $100,000 for individual researchers working on a given topic, or $250,000 for teams. Grant recipients must attend a meeting to discuss their work with colleagues, which offers an extra incentive to get creative. "If a new idea comes out of the meeting, we encourage people to write a two-page application on site — which we'll fund if we think it is possible," says Jim Gentile, president of the RCSA. The foundation launched a Scialog on enhancing solar cells in 2010; another, on energy storage, will be launched this year.

Individual institutions are also promoting innovative approaches. To take advantage of the expertise spread across departments, scientists at the University of Michigan in Ann Arbor sought a "fast, interdisciplinary funding vehicle that doesn't have the downside of peer review", says Thomas Zurbuchen, associate dean for entrepreneurship. They came up with MCubed, a 2-year pilot project funded with $15 million from the provost and individual university schools, colleges and investigators.

University researchers can register with the MCubed website (http://mcubed.umich.edu) and float their ideas to the community. Each is allotted a token for $20,000; to unlock and combine the funding, three researchers from different disciplines have to establish a team and register their project. Once they've done that, they immediately receive their combined $60,000 to hire staff and begin work. The teams must draft a mentoring plan to protect participating students' academic progress and must give a talk about the project after it ends. The website launches this summer, says Zurbuchen, and should fund its first ideas by the end of the year. "We want to swing for the fences, realizing we may have some failures on the way to some massive successes," he adds.

RISK MANAGEMENT

How can applicants endure without losing funds if their risk doesn't pay off? In the NIH's high-risk, high-reward programmes, "if an idea isn't developing the way it was expected to, awardees have the flexibility to pursue a more promising avenue of research", says Basavappa, adding that he cannot recall a requested change in course ever being denied.

ARPA-E takes a different tack, instilling a rigid level of oversight — something some researchers may not like. Instead of grants, the agency uses cooperative research agreements, which pay incrementally for work performed, giving ARPA-E the authority to remove funding if projects don't meet expectations in on-site visits and tangible milestones at decision points every three months. Toone says that about 10% of projects are spiked. "We take on more technical risks and we manage that risk," he says.

High-risk, high-reward research can break down barriers and bring diverse teams together, but some researchers are not cut out for life on the edge. "There is a self-selection of those applicants willing to take a risk," says Basavappa. Alf Game, acting director of research at the BBRSC, agrees: "Not everybody is capable of or wants to be at the cutting edge of every damn thing they are doing." ■

Virginia Gewin *is a freelance writer based in Portland, Oregon.*

GRADUATE STUDENTS

Unionization review

Graduate-student assistants at private US universities may once more be eligible to join a union if a 2004 federal ruling that blocks formation of bargaining units is reversed. On 22 June, the US National Labor Relations Board (NLRB) voted to review the ruling. A solicitation for legal comment closes on 23 July. The 2004 ruling said that graduate students are not employees and cannot elect unions; in doing so, it overturned a 2000 decision. Graduate students at New York University and the Polytechnic Institute of New York University have petitioned the NLRB for an election in the past two years. Nancy Cleeland, director of public affairs at the NLRB, says that no date has yet been set to review the ruling.

TRAINING

Clinical course for PhDs

To broaden career options, the US National Institutes of Health (NIH) has launched a scheme to introduce biomedical PhD students to clinical and translational research. The two-week programme at the NIH Clinical Center in Bethesda, Maryland, began on 9 July. Students will learn principles of clinical and translational research design, implementation and analysis; participate in a mock institutional review board; and learn how to apply for a drug to be approved by the US Food and Drug Administration. "We wanted to open students' eyes to the fact that there are opportunities beyond core, basic research," says Frederick Ognibene, a deputy director at the clinical centre. Next year's programme will incorporate feedback and is expected to include more participants.

EARLY-CAREER RESEARCHERS

Advocacy group forms

A cross-border coalition of researchers has formed to advocate for better working conditions and to inform and inspire policy. The International Consortium of Research Staff Associations (ICoRSA) will address early-career challenges including low wages, limited career prospects, mobility restrictions and inadequate recognition. "The same issues exist in almost every country, and we felt that they have to be addressed globally," says Cathee Johnson Phillips, executive director of the US National Postdoctoral Association, one of ICoRSA's founding members. ICoRSA held its first meeting on 14 July at the 2012 Euroscience Open Forum in Dublin.

WHITE LIES

A helping hand.

BY GRACE TANG

"Anthony, is it normal at our age not to remember parts of our lives? Parts people would consider important?"

I froze for the smallest split second, but years of acting had trained me well. In fact, there were days when I forgot that my colleague was not what he appeared to be. I willed my fork to resume its passage from my mouth back to my plate, slowly and calmly.

"Why do you ask, Darren?"

"I was talking to a student of mine who's graduating soon. He's very excited, naturally."

I nodded as we both gave up pretending to care about lunch.

"Problem was, when I tried to recall my own graduation, I drew a blank."

My heart was racing. Lisa would not be happy to hear this. While he spoke, I typed furiously but stealthily on my phone under the table. *Subject Three is catching on.*

"It gets worse. After more thought, I realized I could recall only the barest details about my time in college."

I maintained my perfect poker face, "Hmm. I guess I don't remember much from college either." Fond memories of college flooded my brain.

My phone buzzed, balanced on my knee. I glanced down. *Come now.*

"Gotta go?" Darren had caught me looking at my phone.

"Uh, yeah, Lisa wants to see me."

He'd noticed my nervousness. "The problem with collaborating with your wife, huh? Never know whether you're in trouble because of work, or because you forgot your anniversary."

Lisa looked much older than her 40 years as I entered her office, out of breath. "What happened?" she asked.

"It was his missing memories of graduation that triggered it."

"Damn, those were always the hardest," she rubbed her fingers on her temples. "It's almost impossible to fake memories of a major life event."

We had been in graduate school together when she'd started work on implanting information directly into the brains of rhesus macaques. Almost like magic, her monkeys knew where food was hidden in rooms they had never been in, and recognized other monkeys they'd never met.

NATURE.COM
Follow Futures on Facebook at:
go.nature.com/mtoodm

When she managed to impart basic mathematics to her charges with no effort on their part, her work was broadcast on every major news network in the world. Lisa should have been the happiest person in the scientific community. Instead, one evening, I found her sitting on the floor in the corner of the lab, face in her hands.

"Lisa, what's wrong?"

She looked up and wiped the smudged mascara from her cheeks.

"The Dean of Research visited me today. He said the world hadn't seen anything this exciting since Dolly the sheep."

"And that's bad because …?"

"Like cloning, it's never going to move past animal work. They won't let me use human subjects."

But I knew it would take more than rules to stop Lisa. When her research assistant, a mediocre student at best, started acing every exam a few months later, I knew exactly what was going on. I still remember the night we were the last two people in lab, and I seized my chance.

"How are you doing it?"

Lisa struggled to contain her smile, as if glad that someone had finally figured it out. She checked to see no one else was around. "It wasn't stable at first … as soon as she realized there was no way she could know all the stuff she did without having ever gone to a single class, the knowledge vanished."

"Looks like it's working now."

"It was an easy fix — I figured out that unlike the macaques, humans couldn't handle the sudden unexplained appearance of vast amounts of factual knowledge. So when I put facts and skills in her brain, I also threw in memories of having gone to lectures, studying, all that stuff."

It was then I realized why the project had been stopped.

"Granted, autobiographical memories are much harder to implant than semantic facts. It's very similar to hypnosis — you suggest something to them, and their brains fill in the rest."

"So in other words, you're telling people very convincing lies?"

"Just white lies, Anthony …"

When I still looked unsure, she led me to her equipment room — she rarely let anyone back there. I was honoured.

"How'd you like to work on the next one with me?"

I slept on it. Half of me wanted to report this to the authorities, but it was too good an opportunity to pass up. And by then, I realized I liked Lisa for more than her intellect …

The first time Lisa brought Darren to the lab, I smelt him before I saw him. Plucked from the streets, he hadn't had a shower in days. And yet five years later, Darren was a fellow assistant professor, about to deliver a lecture on molecular neuroscience down the hall.

Lisa paced in front of me. "I was stupid. I was depending too much on the human mind's ability to fool itself. Just suggest to someone they were abused by their father as a child, and they'll tell you under oath how it happened. Here I am, hard-wiring memories into his head, and he doesn't buy it. What more can I do?"

I couldn't keep it in any longer.

"Lisa, do you ever feel this is wrong?"

"Not this again …" she sighed. "We took a homeless, illiterate man off the streets and made him a genius. How is this wrong?"

Defeated, I left for my office. Work was taking its toll on our relationship. Deep in thought, I fiddled with my wedding ring.

The blood drained from my face. Try as I might, I could not recall a single detail of my wedding day. ■

JACEY

Grace Tang *is a graduate student in psychology at Stanford University. Writing short stories is one of her favourite forms of structured procrastination.*

PRENATAL DIAGNOSTICS

Fetal genes in mother's blood

The genome sequence of a fetus can be inferred from the relative numbers of variants of DNA sequences in a pregnant woman's blood. This advance in non-invasive diagnostics comes with some ramifications. SEE ARTICLE P.320

DIANA W. BIANCHI

Until the mid-twentieth century, medical examination of a human fetus was surprisingly crude, consisting principally of uterine palpation to assess fetal growth. Over the past 35 years, however, progress in fetal-imaging techniques, combined with measurement of the chemical composition of maternal blood, has substantially improved the assessment of fetal health[1]. On page 320 of this issue, Fan *et al.*[2] demonstrate that it is now possible to unambiguously determine the whole genome sequence of a fetus from a teaspoon's worth of maternal blood. The potential repercussions of such non-invasive prenatal screening must be carefully considered, particularly as the development comes at a time when many health-care providers are not familiar with the complex concepts of molecular genetics.

Fan and colleagues determined the entire genetic sequence — the order of DNA nucleotides, represented by the letters A, G, T and C — of two unrelated fetuses by sequencing cell-free DNA circulating in the plasma of their mothers' blood. During pregnancy, cells of the placenta undergo programmed cell death, which continuously releases large amounts of nucleic acids into the maternal bloodstream[3], so that a pregnant woman's blood contains a mixture of her own DNA and that of her fetus. Because humans are diploid, meaning they have two copies of each chromosome, there are three haploid (single-copy) genomes in the woman's circulation: her own two (one that she transmits to the fetus and the other that she does not) and the haploid genome contributed by the father of the fetus (Fig. 1).

The advent of rapid and cost-effective techniques for DNA sequencing means that it is now possible to count the number of specific DNA molecules in a sample. Fan *et al.* designed their experiments based on the recent finding[4,5] that parents transmit long stretches of their DNA as blocks, known as haplotypes, to their offspring. The authors paid particular attention to haplotypes containing sequences in which the two copies of the mother's genome differed at just a single DNA base — a single nucleotide polymorphism. They then counted these differing sequences, and applied the premise that those haplotypes that were numerically over-represented in the plasma DNA would correspond to the maternally-inherited part of the fetal genome (Fig. 1). Moreover, they deduced the paternal contribution by identifying sequences that were present in the plasma DNA but absent from the maternal genome.

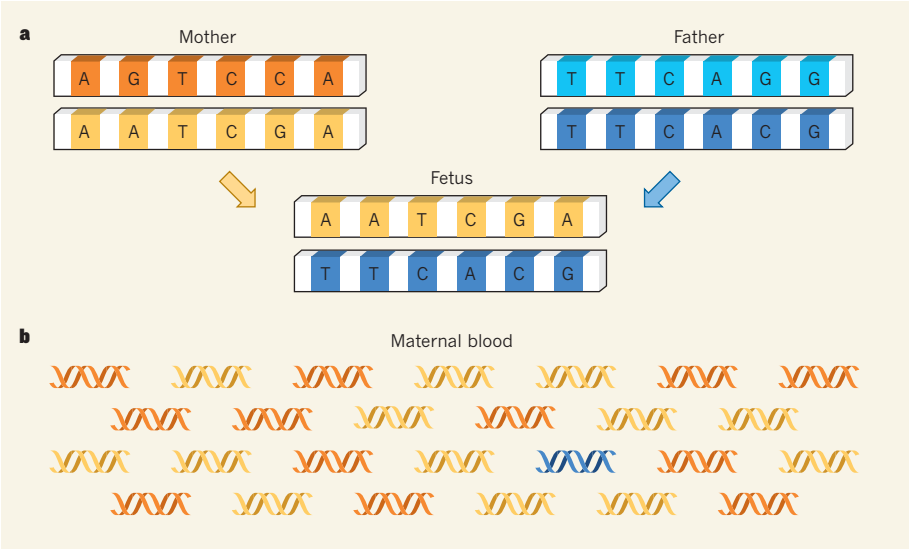


Figure 1 | Deducing a fetal genome. Human cells contain two copies of each chromosome, such that every DNA sequence is represented twice. Egg and sperm cells, however, have only one copy of each chromosome, which comprises a mixture of the sequences from both members of the chromosome pair. Long stretches of DNA sequence, called haplotypes, stay together as a physical unit during the chromosome recombination that occurs during cell replication. When the egg and sperm combine at fertilization, the fetal genome receives one copy of each haplotype from the mother and another from the father. **a**, Small sequence differences between parental chromosomes indicate from which chromosome the haplotype is derived; in this example, the fetal genome at this chromosomal location is derived from the maternal 'yellow' chromosome and the paternal 'dark blue' chromosome. **b**, The blood of a pregnant woman contains both her own DNA and that of her fetus. Therefore, for every haplotype, the blood-derived DNA will contain copies of the paternal sequence inherited by the fetus (blue), copies of the maternal sequence that is not passed to the fetus (orange) and copies of the maternal sequence that was inherited by the fetus (yellow), and which will therefore be present in excess relative to the orange sequence.

In a second set of experiments, Fan *et al.* focused on the more clinically relevant 'exome' portion of the genome — the sequences that encode proteins. The exome is considerably smaller than the entire genome, so the researchers could analyse specific sequences in greater detail and distinguish between sequence variations that were inherited from one of the parents versus those that represented newly occurring mutations.

The authors applied these two approaches to blood samples from two pregnant women. One of the women had a deletion of a large sequence on one of her two copies of chromosome 22, a mutation that is associated with a disorder known as DiGeorge syndrome. By analysing the genes around the deleted region on chromosome 22, the researchers deduced that the haplotype derived from the copy of chromosome 22 containing the deletion was over-represented in the mother's blood, indicating that the fetus shared that section of DNA and was therefore similarly affected by the condition. This finding demonstrates that a non-invasive fetal diagnosis can be made even when the fetus shares the same mutation as its mother.

This advance comes within a month of another report, by Kitzman and colleagues[6], in which fetal genotype was inferred from DNA sequences obtained from blood samples from the mother, father and from umbilical

cord blood. Although having a paternal DNA sample makes such analysis easier, Fan and colleagues' study shows that it is not necessary. Furthermore, comparing the father's DNA sequence with that of the fetus carries the risk of uncovering mistaken paternity, and this is avoided in Fan and colleagues' approach.

How will the ability to non-invasively sequence the fetal genome improve prenatal care? Fan *et al.* posit that it will enable treatment for genetic disorders to begin immediately after delivery. I argue that we could most effectively use the information to begin treatment while the fetus is still in the womb[7]. However, it is striking that before we have even considered all of the ramifications of complete genomic sequencing of a newborn's DNA, we now have three demonstrations of non-invasive sequencing of the fetal genome[2,6,8]. The situation is ethically and clinically more complex with a fetus than with a newborn for two reasons: one, the 'patient' is in the womb and cannot be fully examined, and two, prospective parents have the option of terminating the pregnancy.

These studies therefore raise many ethical and practical questions about how prospective parents and physicians might use this genomic information. For example, Kitzman and colleagues[6] detected 44 spontaneous point mutations in the fetal genome that they sequenced. One of these mutations creates an amino-acid substitution in the protein encoded by the gene *ACMSD*, which is implicated in Parkinson's disease, suggesting that this mutation might have clinical significance later in that unborn child's life. Will expectant couples want to know this sort of information? Now, multiply this point mutation by several hundred — a plausible quantity of 'noteworthy' genetic information that might typically be obtained from a whole-genome sequence — and imagine the time and resources needed to provide parents-to-be with genetic counselling regarding the implications of all of this data.

Although the concept of routine fetal-genome sequencing may still seem futuristic, non-invasive prenatal diagnosis of abnormal chromosome number is already offered to pregnant women in certain high-risk categories in the United States and China[9]. But before the vast amounts of information acquired from fetal-genome sequencing can be applied in a useful manner, the gap between technology and clinical interpretation must be narrowed. For parents to learn their fetal ACGTs, substantial investment is needed in teaching health-care providers about the human genome. ■

Diana W. Bianchi *is at the Mother Infant Research Institute, Tufts Medical Center, and in the Departments of Pediatrics, Obstetrics and Gynecology, Tufts University School of Medicine, Boston, Massachusetts 02111, USA.*
e-mail: dbianchi@tuftsmedicalcenter.org

1. Bianchi, D. W. & Ferguson-Smith, M. A. *Prenat. Diagn.* **30,** 601–604 (2010).
2. Fan, H. C. *et al. Nature* **487,** 320–324 (2012).
3. Lo, Y. M. *et al. Lancet* **350,** 485–487 (1997).
4. Fan, H. C. *et al. Nature Biotechnol.* **29,** 51–57 (2011).
5. Kitzman, J. O. *et al. Nature Biotechnol.* **29,** 59–63 (2011).
6. Kitzman, J. O. *et al. Sci. Transl. Med.* **4,** 137ra76 (2012).
7. Bianchi, D. W. *Nature Med.* **18,** 1041–1051 (2012).
8. Lo, Y. M. *et al. Sci. Transl. Med.* **2,** 61ra91 (2010).
9. Benn, P. *et al. Prenat. Diagn.* **32,** 1–2 (2012).

The author declares competing financial interests.
See go.nature.com/swrest for details.

BIOGEOCHEMISTRY

The great iron dump

The discovery that marine algal blooms deposit organic carbon to the deep ocean answers some — but not all — of the questions about whether fertilizing such blooms is a viable strategy for mitigating climate change. SEE ARTICLE P.313

KEN O. BUESSELER

"Give me half a tanker of iron and I'll give you the next ice age," is perhaps the best-known quote in ocean science. It comes from the late John Martin[1], a leader in the study of iron and its role in sustaining productivity in the ocean. The quip refers to Martin's proposal that the addition of iron to the upper ocean could trigger algal blooms that would ultimately alter climate by sequestering atmospheric carbon dioxide as organic carbon in the deep ocean. Smetacek *et al.*[2] have taken on the challenge of proving Martin's hypothesis experimentally, and on page 313 of this issue they report that carbon formed from iron-fertilized algal blooms does indeed sink to the deep ocean — the first time that this has been convincingly observed.

Productivity in many parts of the global ocean is limited by iron levels, as demonstrated through several studies[3] in which the addition of iron to the upper ocean stimulated phytoplankton blooms and greatly increased CO_2 uptake into surface waters through photosynthesis. But for ocean iron fertilization (OIF) to have an impact on Earth's climate, organic carbon produced by the phytoplankton must be transported to the deep ocean where it cannot readily re-exchange with the atmosphere — this is the key event in Martin's ice-age-inducing scheme. Proving Martin's iron hypothesis therefore requires the fate of blooms to be followed.

This is what Smetacek *et al.* have done as part of the European Iron Fertilization Experiment (EIFEX). By tracking phytoplankton biomass using several methods, the authors demonstrated that at least half of the carbon captured by the algal bloom in their OIF experiment sank to depths well below 1,000 metres, some of which is likely to have reached the sea floor. Their findings help to inform us about how the oceans regulate atmospheric CO_2, and provide further input to the debate into whether the oceans can, or should, be deliberately modified using OIF to mitigate the effects of climate change — an example of a practice known broadly as geoengineering.

OIF experiments are challenging because the waters used in such studies cannot usually

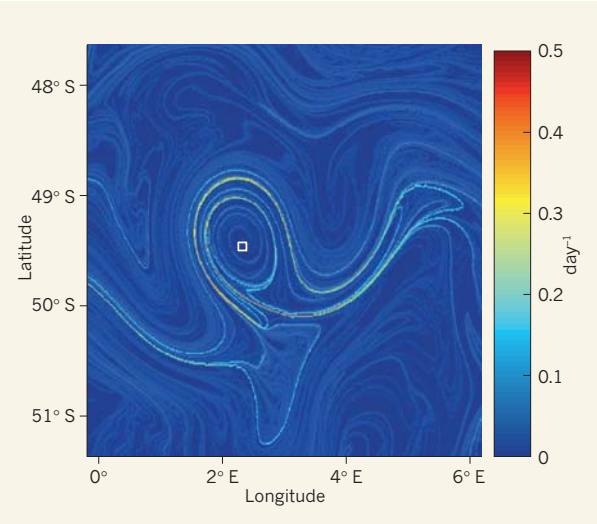


Figure 1 | Ocean eddy. Smetacek *et al.*[2] describe the results of an experiment in which they added iron salts to a patch of ocean within an eddy in the Southern Ocean, near Antarctica. The eddy is depicted here using Lyapunov exponents (reported as day^{-1}). Lines of maxima of Lyapunov exponent represent barriers to the transport of water in the ocean, and can be thought of as fronts between water masses of different origins. The white square corresponds to the centre of the ocean patch to which iron was added. The authors show that algal blooms triggered by the introduction of iron deposit organic carbon to the deep ocean.

be separated from the rest of the ocean in the way that laboratory experiments can be constrained by beakers. To overcome this problem, Smetacek *et al.* used an ocean eddy near Antarctica as a 'beaker' (Fig. 1). This solution seems to work well — the authors provide considerable evidence that the upper and lower layers of the eddy moved together coherently, and that the eddy had exchanged less than 10% of its content with the surrounding ocean by the end of the experiment.

The authors introduced dissolved iron(II) sulphate ($FeSO_4$) over a 167-km^2 patch in the eddy's core, so that the concentration of iron at the ocean's surface reached a level known to stimulate phytoplankton growth. The consequences were substantial: phytoplankton biomass more than doubled in 24 days, with 97% of the observed increase in chlorophyll associated with large diatoms, a class of phytoplankton that has high iron requirements. Along with this growth, the authors observed a reduction in levels of dissolved inorganic carbon (DIC) and of several nutrients (nitrogen, phosphorus and silicon). Data collected from stations outside the eddy, used as controls to monitor non-fertilized conditions, showed no such effects.

The scientists kept up their study for a full 37 days — longer than any other OIF experiment — and so were able to document the collapse of the diatom bloom through the formation of rapidly sinking aggregates of dead phytoplankton and zooplankton faecal pellets that carried carbon to the deep ocean. The last 13 days of observations were crucial to their success, because they enabled the authors to calculate the depletion of dissolved and particulate carbon at the surface and subsequent increases in particulate organic carbon at depth. Such 'budgets' are notoriously tricky to close in OIF studies, because of the difficulty in quantifying carbon losses that occur through air–sea gas exchange and physical mixing at the fertilized patch's boundaries, and because it is hard to account for variability in carbon levels within and outside the patch. In this case, however, the combination of evidence was clear: the iron-induced diatom bloom led to the export and sequestration of about one mole of carbon per square metre of ocean surface, from the uppermost 100 metres of ocean. In fact, one of the methods used by the authors suggested that, at its peak, carbon flux was the largest ever recorded in the Southern Ocean.

The implications of these findings are several-fold. First, a measure of the efficiency of carbon export in the experiments can be obtained by dividing the amount of DIC removed from the upper 100 metres of ocean by the amount of iron added. This measure — the carbon/iron molar ratio — is crucial for geoengineering proposals, which must specify how much iron will be needed to affect climate. In the laboratory, the ratio can be 100,000 or more[4]. By contrast, the ratios reported in previous OIF experiments[3] have been much lower, in part because iron uptake by plankton in the ocean is inefficient compared with that under laboratory conditions, but also because of differences in the amounts of iron and carbon that are recycled at the surface, or which sink to depth. Smetacek *et al.* report that the carbon/iron molar ratio in their long experiment was 13,000 — higher than in the previous OIF studies — and argue that this number would have increased further had they followed the bloom for longer.

Furthermore, the authors' results defied expectations[5] that the availability of light would limit phytoplankton growth in their experiment. Phytoplankton grow in the 'mixed layer' of the ocean, the region in which the uppermost layers of the ocean are homogenized by wind and other physical effects; the mixed layer in Smetacek and colleagues' experiment was deep, extending down to 100 metres, where little light would penetrate. Comparison of Smetacek and colleagues' study with naturally occurring blooms[6,7] in iron-rich waters near islands in the Southern Ocean also suggests that their experiment was similar to natural OIF events, and that higher sequestration was potentially possible.

Although the authors conclude that OIF does indeed sequester carbon in the deep ocean, questions remain about the possible unintended consequences of geoengineering. For example, OIF might cause undesirable effects, such as the production of nitrous oxide (a more potent greenhouse gas than carbon dioxide); oxygen depletion in mid-waters as algae decompose; or stimulation of a toxic algal bloom. And, as with all carbon-removal methods, OIF is no silver bullet for mitigating climate change. The ocean's capacity for carbon sequestration in low-iron regions is just a fraction of anthropogenic CO_2 emissions, and such sequestration is not permanent — it lasts only for decades to centuries. However, humans have already embarked on an ocean geoengineering experiment through our energy practices (which are affecting climate and acidifying the seas), by fishing, and through our other uses of ocean resources.

Most scientists would agree that we are nowhere near the point of recommending OIF as a geoengineering tool. But many think[8,9] that larger and longer OIF experiments should be performed to help us to decide which, if any, of the many geoengineering options at hand should be deployed. EIFEX certainly does not answer all of the questions about geoengineering, but by showing how the addition of iron to the ocean not only enhances ocean productivity, but also sequesters carbon, it is one of the best OIF studies so far. ■

Ken O. Buesseler *is in the Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA.*
e-mail: kbuesseler@whoi.edu

1. Martin, J. H. in *US Joint Global Ocean Flux Study Newsletter* **1** (2), (US JGOFS Planning Office, Woods Hole Oceanographic Institution, 1990).
2. Smetacek, V. *et al. Nature* **487,** 313–319 (2012).
3. Boyd, P. W. *et al. Science* **315,** 612–617 (2007).
4. Sunda, W. G. & Huntsman, S. A. *Mar. Chem.* **50,** 189–206 (1995).
5. de Baar, H. J. W. *et al. J. Geophys. Res.* **110,** C09S16 (2005).
6. Blain, S. *et al. Nature* **446,** 1070–1074 (2007).
7. Pollard, R. T. *et al. Nature* **457,** 577–580 (2009).
8. Buesseler, K. O. *et al. Science* **319,** 162 (2008).
9. http://isisconsortium.org/

CARDIOLOGY

Bad matters made worse

Heart attacks occur when lipoprotein-driven inflammation called atherosclerosis triggers blood clotting in the arteries. It seems that the attacks can, in turn, accelerate atherosclerosis by fanning the inflammation. SEE LETTER P.325

IRA TABAS

Heart attack, or myocardial infarction, is a leading cause of morbidity and mortality worldwide, and people who have had one infarction are at increased risk of another in the first year or so after the attack[1]. Myocardial infarction results from acute, occlusive thrombosis (blood clots) within the coronary arteries. These clots form at sites of atherosclerosis, a chronic disease process in which fat and cholesterol build up along the artery walls[2]. Atherosclerosis starts when circulating fat-carrying particles called lipoproteins, most notably low-density lipoprotein (LDL), are retained in the subendothelium, a tissue layer in the artery wall[3]. This induces an inflammatory response that involves the influx of immune cells called monocytes, which differentiate into other inflammatory-cell types, including phagocytic cells called macrophages and dendritic cells[3,4]. On page

325 of this issue, Dutta *et al.*[5] show that in mice with atherosclerosis, myocardial infarction leads to increased monocyte recruitment and enhanced atherosclerosis. If these processes are similarly linked in humans, the findings may have implications for therapeutic strategies in human heart disease.

The idea for this study came from the observation that a high monocyte count in the blood after myocardial infarction is a risk factor for repeat infarction, and that, in mice, infiltrating spleen-derived monocytes have a role in atherosclerosis[6]. If there is a cause–effect relationship between monocyte count and atherosclerosis, this would suggest that the number of circulating monocytes, particularly monocyte subclasses that are especially inflammatory, could be a rate-limiting factor.

To explore these issues, Dutta and colleagues used a mouse model of atherosclerosis in which the mice lack apolipoprotein E (APOE), a protein that facilitates the removal of certain types of atherosclerosis-promoting lipoproteins from the blood. The mice were also fed a high-cholesterol diet. Mice do not normally develop atherosclerosis because they have low levels of atherogenic lipoproteins, but the combination of a high-cholesterol diet and the absence of APOE in this model induces atherosclerosis. However, even this robust model does not cause acute thrombosis and myocardial infarction, probably owing to several physical and biochemical factors. So the authors modelled heart attack in the atherosclerotic mice by clamping shut their left coronary artery.

Within 1–3 weeks after myocardial infarction, the atherosclerotic lesions in the aortas of the 'heart attack' mice were approximately 40% larger than those in sham-operated animals, which had atherosclerosis but had not undergone infarction. The mice with infarction also had elevated blood monocyte counts, and the lesions themselves contained greater numbers of inflammatory cells and showed signs of disease progression, including larger regions of dead cells (necrosis). Furthermore, the authors provide evidence that the sympathetic nervous system (SNS) was activated in the mice following myocardial infarction, and that this led to an expansion of monocytes in their spleens. This was followed by monocyte delivery to the blood (a process called monocytosis) and then to the atherosclerotic lesions (Fig. 1). The SNS is associated with the 'fight or flight response', so it may be that in humans, and perhaps in this mouse model, SNS activation following myocardial infarction is a response to pain, anxiety and an acute decrease in heart function. The authors also showed that splenectomy or drug-induced blockade of the SNS lowered post-infarction monocytosis in the mice, although they did not report whether these interventions affected progression of atherosclerosis per se.

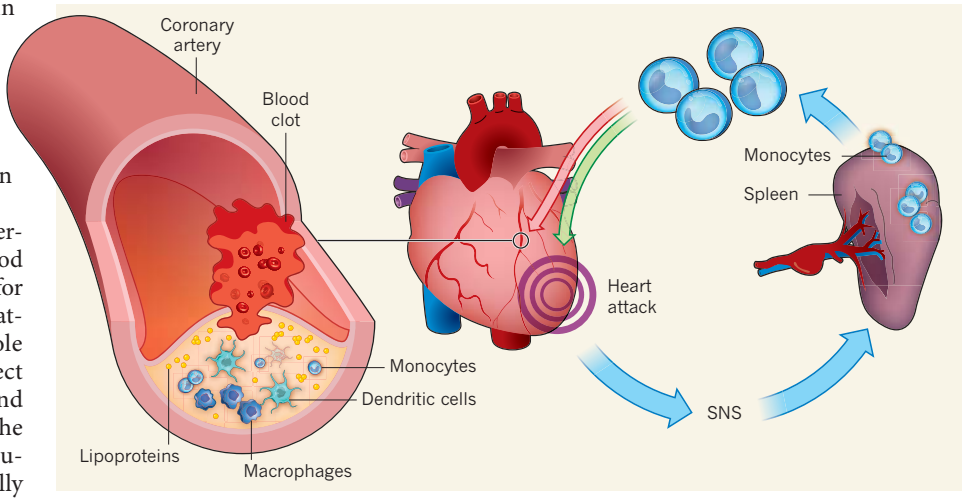


Figure 1 | A cycle of damage and repair. During atherosclerosis, fat-carrying lipoproteins are retained in the artery wall. This induces an inflammatory response that is characterized by an influx of immune cells called monocytes, which differentiate into other inflammatory cells, such as macrophages and dendritic cells. Blood clots at sites of atherosclerosis can block the arteries and cause heart attacks. In a mouse model of atherosclerosis, Dutta *et al.*[5] show that heart attack activates the sympathetic nervous system (SNS), and that this promotes monocyte production in the spleen. The authors demonstrate that the monocytes move from the spleen to the blood and then to atherosclerotic lesions in the arteries, where they accelerate the progression of these lesions (red arrow). However, monocytes are also likely to be important in the repair of heart muscle tissue (green arrow) following heart attack[9], and this dual role could complicate attempts to treat post-heart-attack atherosclerosis by targeting monocytes.

Might this process of heightened susceptibility following myocardial infarction also be important in humans? The mouse model used by Dutta *et al.*[5] has many obvious differences to human disease, most notably the absence of atherosclerosis-driven acute thrombosis or infarction, and the possible confounding effects of the absence of APOE, which has several other biological actions besides lipoprotein clearance. However, the atherogenic process itself is similar to that which occurs in humans, and the links to post-infarction monocytosis and SNS activation in humans add relevance.

Even so, the mechanism underlying post-infarction susceptibility in humans is likely to be multifactorial. For example, heightened systemic and cardiac inflammation could fuel the inflammatory response by mechanisms over and above raising monocyte numbers, and biological changes induced by the occluded artery or by tissue damage sustained in the attack might alter the subendothelium of nearby arteries in ways that promote lipoprotein retention. It should also be remembered that heart attacks often indicate the presence of environmental and/or genetic risk factors, and so are probably a marker of individuals predisposed to accelerated atherosclerosis due to a variety of factors beyond the actual infarction event.

The possibility that an SNS–spleen–monocyte pathway of accelerated atherosclerosis contributes to the increased risk of repeat events after heart attack could add perspective to current therapeutic strategies and suggest new ones. Several clinical trials have shown that post-infarction administration of high doses of statins — drugs that lower blood levels of LDL — have a substantial protective effect against repeat infarction. Reduction of LDL would be expected to interrupt the atherosclerotic process in the post-infarction period, as it does in other settings. However, some researchers interpret this rapid effect of statins as indicative of other mechanisms, such as anti-inflammatory activity, which is known to be a property of statins[7]. Low levels of high-density lipoprotein (HDL), a risk factor for myocardial infarction, have also been linked to increased monocytosis and atherosclerosis[8], so the current study could inform current human therapeutic trials designed to raise plasma HDL. It would also be interesting to assess whether beta-blocker drugs, which are often given to patients following heart attack, have any effect on monocytosis and atherosclerosis progression.

Finally, although therapeutic strategies aimed at lowering monocyte numbers may have broad benefit in blocking the progression of atherosclerosis, their applicability could be tempered by the fact that post-infarction monocytosis could also have a role in healing the damaged heart muscle[9] (Fig. 1). Indeed, this previous finding, in the context of Dutta and colleagues' study, raises the fascinating question of why these monocytes help tissue healing in infarcted heart muscle but worsen atherosclerosis. A crucial healing function of monocyte-derived phagocytes is in the clearance of dead cells[10]. It is possible that dead cells in atherosclerotic lesions are less easily recognized by the newly arriving phagocytes than are dead cardiac muscle cells. Alternatively, when the post-infarction phagocytic cells enter the lesions, they may be exposed

to factors that impair their ability to clear dead cells — factors that would not be present in the infarcted heart muscle. Thus, although it is likely that inflammation and monocytosis have several roles in the response to heart attack, further understanding of these processes is needed to intelligently translate these concepts into new therapies. ■

Ira Tabas *is in the Departments of Medicine, Pathology and Cell Biology, and Physiology and Cellular Biophysics, Columbia University, New York, New York 10032, USA.*
e-mail: iat1@columbia.edu

1. Milonas, C. *et al. Am. J. Cardiol.* **105,** 1229–1234 (2010).
2. Libby, P., Ridker, P. M. & Hansson, G. K. *Nature* **473,** 317–325 (2011).
3. Williams, K. J. & Tabas, I. *Arterioscler. Thromb. Vasc. Biol.* **15,** 551–561 (1995).
4. Moore, K. J. & Tabas, I. *Cell* **145,** 341–355 (2011).
5. Dutta, P. *et al. Nature* **487,** 325–329 (2012).
6. Robbins, C. S. *et al. Circulation* **125,** 364–374 (2012).
7. Bu, D. X., Griffin, G. & Lichtman, A. H. *Curr. Opin. Lipidol.* **22,** 165–170 (2011).
8. Yvan-Charvet, L. *et al. Science* **328,** 1689–1693 (2010).
9. Leuschner, F. *et al. J. Exp. Med.* **209,** 123–137 (2012).
10. Tabas, I. *Nature Rev. Immunol.* **10,** 36–46 (2010).

CHEMICAL BIOLOGY

Greasy tags for protein removal

Most proteins in the human body are difficult targets for small-molecule drugs. This problem may have been overcome with the discovery of molecules that induce protein degradation, suggesting fresh, modular approaches to drug discovery.

TAAVI K. NEKLESA & CRAIG M. CREWS

It was recently discovered[1,2] that proteins covalently 'tagged' with small, synthetic, hydrophobic molecules are degraded by the cell's quality-control machinery. Writing in *Chemistry & Biology*, Long *et al.*[3] now report that non-covalent binding of such molecules also marks proteins for degradation. This finding could open up a wide range of proteins as targets for drug-discovery programmes.

The dearth of newly approved drugs in the past decade reflects the challenges faced by the pharmaceutical industry. Although advances in genomics have identified many proteins that are implicated in disease, many of these proteins — especially those that are not enzymes — are not currently viable drug targets. In fact, it has been estimated that only about 15% of the human proteome is 'druggable' with small molecules[4].

Many attractive drug targets have therefore been dubbed 'undruggable'. For instance, there are roughly 1,400 human transcription factors — proteins that regulate messenger RNA synthesis from DNA, but which lack enzymatic activity. These proteins remain largely undruggable, despite the fact that aberrant expression of some of them is known to cause cancer. One possible solution to this challenge has been the development of small interfering RNAs (siRNAs), which intervene in gene expression by binding to mRNA. However, delivering siRNAs to their targets *in vivo* has been a difficult hurdle to overcome, and so small molecules that can affect the function of undruggable proteins are needed.

Another emerging approach is to destroy, rather than inhibit, target proteins in cells. Normal protein turnover in cells is mainly mediated by the ubiquitin–proteasome system (UPS), which tags unwanted or misfolded proteins with chains of the ubiquitin protein. Once ubiquitinated, the marked proteins are recognized by the proteasome, a large, barrel-like molecular machine that cleaves proteins into small peptides. Efficient removal of unwanted proteins is key to cell survival, as evidenced by the development of proteasome inhibitors as effective antitumour agents[5].

Several strategies have been reported that co-opt the UPS for targeted protein degradation. One of these uses 'proteolysis-targeting chimaeric molecules' to bring the protein of interest close to a ubiquitin ligase (an enzyme that mediates the ubiquitination of a target protein), thus bringing about protein ubiquitination and subsequent degradation[6].

An alternative approach is to mimic a misfolded protein state using small molecules. Normally, the 'greasy' (hydrophobic) side chains of polypeptides are buried in the interior of a globular protein, with the hydrophilic amino-acid residues lying at the surface. Even a small increase in surface hydrophobicity can make a protein unstable. For instance, the deletion of a single amino acid from the CFTR protein is the main cause of cystic fibrosis. The deletion results in the exposure of hydrophobic patches on the surface of CFTR, leading to misfolding and subsequent degradation of the protein (Fig. 1).

We have recently shown[1,2] that the covalent attachment of a synthetic hydrophobic group (such as adamantane, a bulky hydrocarbon) to the surface of proteins attracts chaperone proteins whose job it is to help refold misfolded proteins, or, if they cannot be refolded, to target them for degradation by the proteasome. But most drugs bind to proteins through non-covalent interactions, and it was unclear whether non-covalently bound molecules could also trigger this sequence of events.

Long *et al.*[3] have settled this concern. They investigated the biological effect of attaching a hydrophobic group (Boc_3Arg, a modified arginine amino acid) to trimethoprim (TMP), a ligand molecule that binds non-covalently to the dihydrofolate reductase (DHFR) enzyme from the bacterium *Escherichia coli*. The authors observed that TMP–Boc_3Arg induces 30–80% DHFR degradation in mammalian cells, depending on the rate of DHFR synthesis. This effect could be blocked either by TMP, which competes with TMP–Boc_3Arg for binding to DHFR, or by inhibitors of proteasome activity.

The authors also demonstrated that the glutathione S-transferase (GST) enzyme is degraded when treated with a compound in which Boc_3Arg is attached to ethacrynic acid (EA), a GST inhibitor that becomes covalently bound to the enzyme's active site. This demonstrates that the degradation effect of Boc_3Arg occurs for at least two enzymes. Long *et al.* went on to make a fusion protein in which DHFR is attached to GST, and then treated cells producing the protein with either TMP–Boc_3Arg or EA–Boc_3Arg. They observed that DHFR–GST was degraded more efficiently by EA–Boc_3Arg, which binds covalently to the protein, than by TMP–Boc_3Arg, which binds non-covalently. This suggests that the covalent attachment of hydrophobic tags to enzymes is the more effective strategy for protein degradation.

As TMP is a high-affinity inhibitor of *E. coli* DHFR, further studies are needed to determine whether a small molecule that is both a protein inhibitor and a degradation signal is more effective in abrogating protein function than a simple inhibitor. As pointed out by the authors, the case of botulinum toxin illustrates the advantage of the degradation approach. The most potent form of this toxin, which causes muscle paralysis, has a half-life in the body of about 3 months. Although an inhibitor of the toxin would be able to suppress toxicity in the short term, elimination of the toxin is obviously a preferable therapeutic approach.

However, the Boc_3Arg moiety is large (almost 500 daltons in mass), and large molecules often have poor pharmacokinetic properties that limit their use as drugs. So, appending it to an existing inhibitor could potentially worsen that inhibitor's pharmacokinetic properties. Curiously, even though TMP has high affinity for *E. coli* DHFR and is thought to have excellent cell permeability,

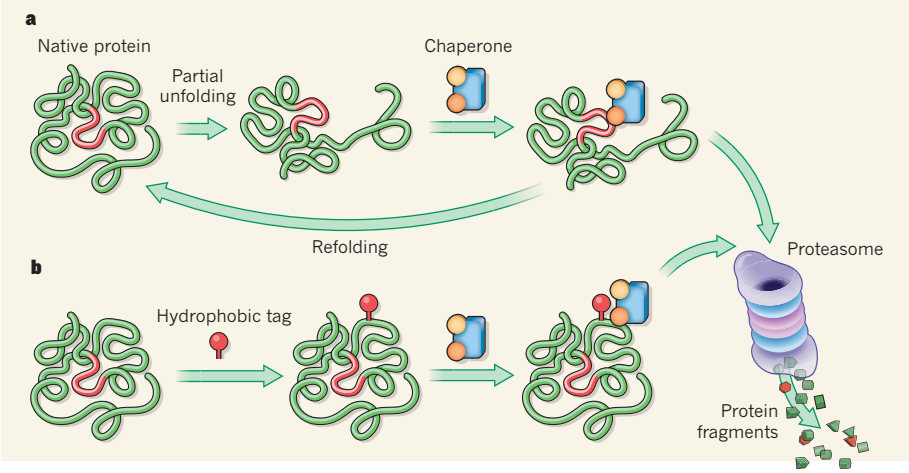


Figure 1 | Hydrophobic tags for protein degradation. **a**, Cellular chaperone proteins help other proteins that have become partially unfolded to refold into their correct tertiary structure. If refolding fails, the chaperones trigger degradation of the unfolded protein by the proteasome, a large protein complex. **b**, Synthetic hydrophobic groups attached to a protein's surface can mimic the partially unfolded state. Because chaperones are unable to refold these proteins, the tagged proteins are degraded by the proteasome. Long *et al.*[3] report that hydrophobic tags do not need to be covalently attached to a protein to induce degradation.

Long *et al.* needed to use a high concentration of TMP–Boc_3Arg to observe protein degradation. This suggests that TMP–Boc_3Arg has difficulty permeating cells.

Other non-covalent ligand–protein systems need to be tested to establish the minimum ligand–protein affinity necessary to initiate protein degradation. Meanwhile, it is intriguing to speculate about how the modularity of this protein-degradation strategy might be used for drug discovery. One could envisage a streamlined process in which ligands for an undruggable protein are identified, appended with hydrophobic moieties (such as adamantane or Boc_3Arg) and tested for their ability to degrade the target protein. Finding high-affinity ligands for undruggable proteins will certainly be a challenge, but methods are becoming available to facilitate this.

For instance, chemical libraries in which each compound is attached to a unique DNA 'barcode' can be tested for protein binding, and the chemical entities that have the highest binding affinities subsequently identified using the barcodes[7]. This method would allow the rapid screening of up to 10^9 compounds, whereas the largest screens currently used assay only about 10^6 compounds[8]. A combination of such high-throughput screening methods with the hydrophobic tagging approach could make today's undruggable proteins attractive biological targets in the search for compounds that ameliorate human disease. ■

Taavi K. Neklesa *and* **Craig M. Crews** *are in the Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA.* **C.M.C.** *is also in the Departments of Chemistry and of Pharmacology, Yale University.*
e-mail: craig.crews@yale.edu

1. Neklesa, T. K. *et al. Nature Chem. Biol.* **7,** 538–543 (2011).
2. Tae, H. S. *et al. ChemBioChem* **13,** 538–541 (2012).
3. Long, M. J. C., Gollapalli, D. R. & Hedstrom, L. *Chem. Biol.* **19,** 629–637 (2012).
4. Russ, A. P. & Lampel, S. *Drug Discov. Today* **10,** 1607–1610 (2005).
5. Kauffman, M. G., Molineaux, C. J., Kirk, C. J. & Crews, C. M. in *Cancer: Principles & Practice of Oncology* (eds DeVita, V. T. Jr, Lawrence, T. S. & Rosenberg, S. A.) 441–449 (Lippincott Williams & Wilkins, 2011).
6. Schneekloth, J. S. Jr *et al. J. Am. Chem. Soc.* **126,** 3748–3754 (2004).
7. Kleiner, R. E., Dumelin, C. E. & Liu, D. R. *Chem. Soc. Rev.* **40,** 5707–5717 (2011).
8. Clark, M. A. *et al. Nature Chem. Biol.* **5,** 647–654 (2009).

NUCLEAR PHYSICS

Nucleons come together

Certain light nuclei can be described in terms of crystal-like arrangements of α-particles, which consist of two protons and two neutrons. The nature of the strong interaction within nuclei may explain such structures. SEE LETTER P.341

MARTIN FREER

An insight into the mechanisms that drive nucleons — protons and neutrons — to form clusters is provided by Ebran and colleagues[1] on page 341 of this issue. They show that a significant driver of this process is the nature of the nuclear potential that confines the neutrons and protons to the nucleus: the deeper the potential, the more defined the clusters become.

The textbook view of the nucleus is a quasi-homogeneous collection of protons and neutrons, which adopt an approximately spherical configuration — a spherical droplet of nuclear matter. However, even in the earliest days of nuclear physics, there was speculation that for light nuclei, the nucleons — which are fermions (particles that have half-integer spin) — might actually arrange themselves into clusters of bosonic character, which have integer spin. This speculation grew out of the experimental observation that the bosonic α-particle (a helium-4 nucleus) was much more stable than other light nuclei. The α-particle, which is composed of two protons and two neutrons, is not only very stable when measured against the energy required to decompose it into its constituent parts, but also extremely inert. The high binding energy of the α-particle could, in principle, make it energetically favourable for nucleon quartets (comprising two protons and two neutrons) to form within the nucleus.

In 1938, Hafstad and Teller[2] conjectured that nuclei such as beryllium-8, carbon-12, oxygen-16 and neon-20 — which consist of an equal, as well as even, number of protons and neutrons — could be described in terms of geometric arrangements of α-particles known as α-clusters. Remarkably, experimental studies show this to be the case — but not quite as imagined. Hafstad and Teller associated the ground states of such nuclei with the α-clusters. However, it is now known that, for the most part, clustering appears in excited states. This can be understood if we consider that, in such light nuclei, excitation energy has to be provided in order to permit the nucleus to undergo α-decay, which generates a daughter nucleus and an α-particle; heavy nuclei can spontaneously experience α-decay. The threshold energy necessary for α-decay corresponds to the difference between the mass of the parent nucleus and the combined mass of the daughter and α-particle, and is precisely the same energy that is required to pre-form the α-particle inside the parent nucleus. This understanding was developed by Ikeda and co-workers[3] in the 1960s.

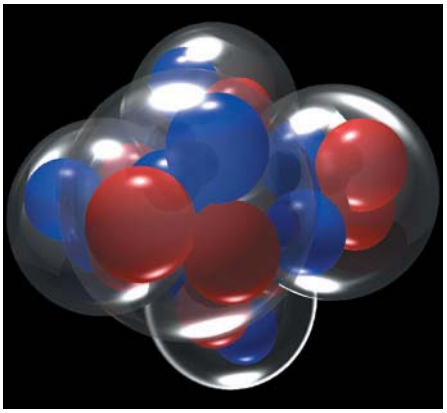
As explained by Ebran and collaborators, these cluster structures have a crucial role in the synthesis of elements in stars. During a star's red-giant phase, energy is generated through the fusion of helium into carbon by the triple-α process. In the first stage of this

process, two α-particles fuse to form ^{8}Be — this nucleus has a cluster structure made up of two α-particles and lives for about 10^{-16} seconds. If, before it disintegrates into two α-particles, ^{8}Be captures a third, ^{12}C is formed. This capture proceeds through a state in ^{12}C known as the Hoyle state, after Fred Hoyle, who proposed its existence[4]. The Hoyle state is the main portal through which ^{12}C is synthesized in nature, and has a pronounced three-α-cluster structure — without the Hoyle state there would be little carbon and no organic life.

Clusters may be considered to be a crystalline phase of nuclear matter because they are a geometric arrangement of the bosonic quartets. How these clusters emerge from the individual fermionic nucleons, which, when not in the cluster form, can be thought of as a fermionic liquid, is an open question.

In their study, Ebran *et al.* assume that the nucleons move such that there is an average force between a given nucleon and all of the other constituents. This force may be associated with an average nuclear potential, which is sometimes called a mean-field potential. The nucleons behave as independent particles and their motion satisfies the Schrödinger equation for this average potential. The quantum energy levels associated with the potential are then filled by the individual protons and neutrons. It is well known[5] that if the potential is deformed, corresponding to a non-spherical nucleus, symmetries in the solution of the Schrödinger equation emerge. The authors discuss how these symmetries correspond to the observed crystalline arrangements of the α-particles. They argue that deformation and the associated symmetries are pivotal drivers of the emergent clustering behaviour.

The main focus of Ebran and colleagues' work reveals that there is a direct link between the depth of the potential (and hence the nature of the nuclear force) and how strongly the symmetries are defined, promoting the formation of the clusters. They use various mathematical entities known as energy-density functionals to generate the nuclear potential for ^{20}Ne. The functionals attempt to build the potential in which the nucleons move by modelling the distributions of nucleon current and density associated with the average nucleon–nucleon interaction. Such an approach is widely used to model the properties of nuclei from the lightest to the heaviest. The authors find a correlation between the depth of the potential and the enhancement of the symmetries of the clustering, as illustrated in ^{20}Ne (Fig. 1). This can be traced to a rather simple factor: the deeper the potential, the shorter the wavelength of the standing waves — which are associated with quantum states — in the centre of the potential. The waves become more 'peaky', enhancing the localization of the nucleons and encouraging them to cluster.



M. FREER

Figure 1 | Clustering in neon. Ebran *et al.*[1] examine the symmetries of the geometric arrangement of α-particles (translucent spheres) that emerges in ^{20}Ne. Each α-particle is composed of two protons (red) and two neutrons (blue).

So, is this the complete picture of nuclear clustering? Although the depth of the potential may emphasize the cluster symmetries, it does not describe the emergence of clustering close to the energy threshold for α-particle decay — the Ikeda picture[3]. There is, therefore, a missing component in this explanation of clustering. Weakly bound nuclei close to the threshold for α-decay can be thought of as open quantum systems, in which the properties of the systems' unbound states (called the continuum) influence, or couple to, the bound states below the α-decay threshold[6]. A complete description of the formation of clustering is challenging, but will need to include such an effect together with the impact of the depth of the potential.

The transformation from the fermionic liquid to the bosonic crystal-like cluster structures reveals key features of the strong nucleon–nucleon interaction within nuclei, and the current work is a step forward in our understanding of this interaction. ■

Martin Freer *is in the School of Physics and Astronomy, University of Birmingham, Birmingham B15 2TT, UK.*
e-mail: m.freer@bham.ac.uk

1. Ebran, J.-P., Khan, E., Nikšić, T. & Vretenar, D. *Nature* **487,** 341–344 (2012).
2. Hafstad, L. R. & Teller, E. *Phys. Rev.* **54,** 681–692 (1938).
3. Ikeda, K., Tagikawa, N. & Horiuchi, H. *Prog. Theor. Phys.* **E68,** 464–475 (1968).
4. Hoyle, F. *Astrophys. J. Suppl.* **1,** 121–146 (1954).
5. Freer, M. *Rep. Prog. Phys.* **70,** 2149–2210 (2007).
6. Okołowicz, J., Płoszajczak, M. & Nazarewicz, W. Preprint at http://arxiv.org/abs/1202.6290v1 (2012).

EVOLUTIONARY PHYSIOLOGY

A bone for all seasons

Because mammals have such high metabolic rates, it has long been thought that their growth is invulnerable to seasonal variation. But their bones turn out to contain annual lines, just as those of cold-blooded animals do. SEE LETTER P.358

KEVIN PADIAN

The bones of at least some animals contain a record of yearly growth, akin to the rings of tree trunks. But what causes this, and does it occur in all animals? These questions arise from our knowledge that corals and molluscs lay down regular growth lines in their skeletons, and that annual growth lines have also been recorded in fishes, amphibians and some reptiles. Because these are slow-growing, cold-blooded animals, the lines were widely assumed[1] to represent pauses in growth resulting from the animals' inability to withstand cyclical environmental stresses such as cold or lack of nutrients. The discovery of these growth lines in dinosaurs[2] fostered the hypothesis that they, too, were slow-growing and had 'typical reptilian' metabolic patterns. But reports[2] of growth lines in some large fossilized birds, as well as in some fossilized and living mammals, considerably weakened this hypothesis.

On page 358 of this issue, Köhler *et al.*[3] present a survey the bones of ruminant mammals hailing from the poles to the tropics, in wet to dry climates, and consistently find growth lines (Fig. 1a)*. Their study contributes to a growing body of literature that undermines the idea that depositing growth lines is a sign of metabolic inferiority.

Because experimental studies of physiology can be performed only on living animals, the metabolic patterns of extinct species are impossible to assess directly. As a result, we tend to think of animals' physiological features in terms of dichotomies — 'cold'- versus 'warm'-blooded, for example — rather than as continua. Furthermore, we often, for convenience, allot extinct forms to one or other of these alternatives on the basis of a single or a few features that seem to correlate well with what we see in the living world[4]. The presence of what are presumed to be annual growth lines in dinosaurs (Fig. 1b) and other extinct reptiles seemed to match the pattern in living

*This article and the paper under discussion[3] were published online on 27 June 2012.

cold-blooded forms. Therefore, dinosaurs were for many decades widely considered to have been cold-blooded.

The problem with this perfunctory generalization was that the bones of warm-blooded extant animals, such as birds and mammals, had never been properly assessed. The bone characteristics of small birds and mammals, which predominate among living forms, were well known. But these species complete skeletal growth in a few weeks or months, so their bones contain no annual growth lines. It was thus presumed that, being warm-blooded, they were unaffected by seasonal vicissitudes, whereas cold-blooded animals always showed annual lines[2].

To the contrary, Köhler *et al.*[3] found that all of the 41 ruminant species they studied — including antelopes, deer and giraffes — deposit growth lines. They show that these are formed annually during the unfavourable season, when the animals lower their body temperatures and metabolic rates, presumably as a way to conserve energy. When the favourable season begins, body temperatures and metabolic rates increase again, and so do growth rates. So it seems that mammals are no different from other vertebrates in this respect.

Do annual growth lines always reflect environmental stress? To explore this question, one group of researchers kept a colony of pygmy lemurs under constant conditions of food and temperature, but varied the light regime to reflect annual periodicity[5]. They then took a subset of the animals and changed the light regime to a 10-month cycle. The bones of both groups deposited growth lines, but the subset did so every 10 instead of every 12 months, falsifying the hypothesis that stress was involved, and instead suggesting the influence of an internal response to light cues, perhaps mediated by the pineal gland in the brain. So, whereas the rhythms of annual growth cycles in vertebrates may originally have reflected environmental stress, it seems that these rhythms have become ingrained in the genes, even in the absence of stress.

What do these findings mean for dinosaurs, and for the interpretation of vertebrate growth in general? In recent years, several studies have chipped away at the cold-blooded dinosaur model. For one thing, the density of blood vessels (vascular canals) in their bones was very high, more comparable to that seen in mammals (Fig. 1) and birds than in reptiles and amphibians[6]. High vascular density implies high blood flow and rapid growth, which can be sustained only by high metabolic rates[2]. The larger dinosaurs and their relatives grew faster than smaller species, a pattern consistent with other vertebrate groups[7]. But young dinosaurs of larger species may sometimes have grown too quickly to leave annual lines in their first year, and other very large ones (such as sauropod dinosaurs) mostly grew too fast to deposit annual lines at all. It seems that these were anything but typical reptiles, and Köhler and colleagues' findings remove another false correlation from this picture.

Another part of this puzzle is the conundrum of determinate versus indeterminate growth. Most warm-blooded animals grow to a typical adult size, at which point growth ceases or radically slows until death; this is called determinate growth. As growth slows, the annual lines appear much closer together, because less bone tissue is deposited between them. Such animals tend to finish long-bone deposition with a layer of tissue that lacks blood vessels and bone cells, in contrast to tissue that is deposited earlier in the growth process. This finishing layer is called the external fundamental system (EFS; Fig. 1), and has not been found in most studies of cold-blooded animals. However, an EFS has been identified in both extant crocodylians[8] and some of their extinct Triassic relatives[9], both of which were previously thought to grow in an indeterminate manner — that is, slowly but continually throughout life. Furthermore, dinosaurs also deposited an EFS when fully grown[10]. Thus, it now seems likely that all vertebrates, living and extinct, have determinate growth, and that we don't see an EFS in slow-growing creatures because most individuals die before they reach full size.

There is still much to learn about the distribution of growth lines in living and extinct vertebrates, as well as what they mean physiologically and how different metabolic regimes of growth and deposition have evolved. But Köhler and colleagues have provided a crucial piece of this puzzle, by conclusively demonstrating that growth lines are typical of large mammals, and thereby discounting the idea that this feature is a characteristic of cold-blooded animals. ■

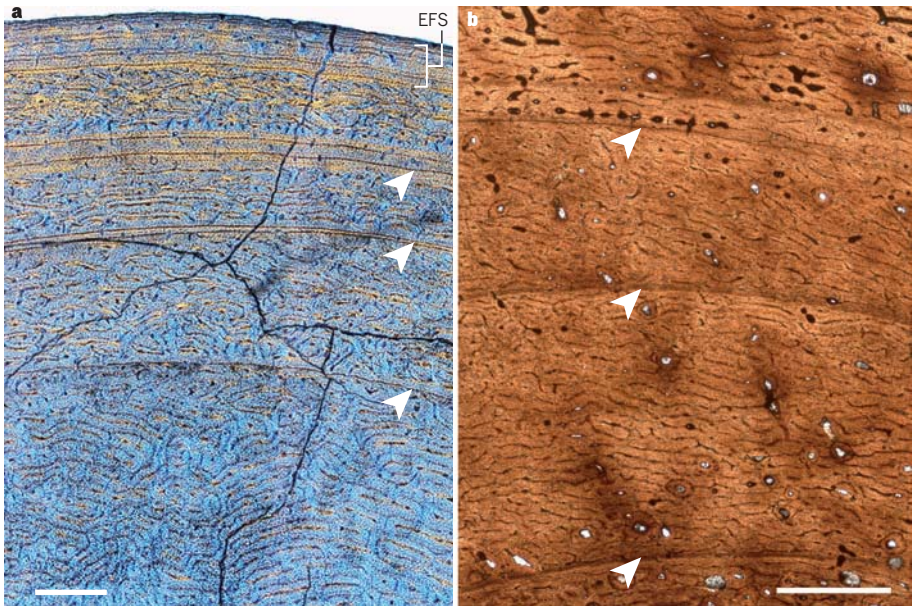


Figure 1 | Marking time. **a**, Köhler *et al.*[3] show that the bones of ruminants contain lines that reflect annual growth patterns, as seen in this cross-section of a long bone from a red deer (*Cervus elaphus*). The arrows point to annual lines that indicate periods during which growth effectively ceased temporarily. The maze-like network of blood vessels appears rather chaotic in the earlier growth stages (lower part of figure), and settles into a more layered pattern later on (upper part of figure). The completion of growth is indicated by a layer of tissue, called the external fundamental system (EFS) that lacks bone cells or blood vessels. Scale: 500 µm. Picture from ref. 3. **b**, Identical features are seen in typical dinosaur fossil bones[2,3,10], such as that of the ornithopod dinosaur *Tenontosaurus tillettii*, shown here. Scale: 1,000 µm. Picture provided by Sarah Werning (modified from ref. 11).

Kevin Padian *is in the Department of Integrative Biology and Museum of Paleontology, University of California, Berkeley, Berkeley, California 94720, USA.*
e-mail: kpadian@berkeley.edu

1. Chinsamy, A. & Hillenius, W. J. in *The Dinosauria 2nd Edn* (eds Weishampel, D. B., Dodson, P. & Osmólska, H.) 643–659 (Univ. California Press, 2004).
2. Padian, K. & Horner, J. R. in *The Dinosauria 2nd Edn* (eds Weishampel, D. B., Dodson, P. & Osmólska, H.) 660–671 (Univ. California Press, 2004).
3. Köhler, M., Marín-Moratalla, N., Jordana, X. & Aanes, R. *Nature* **487,** 358–361 (2012).
4. Bennett, A. F. & Ruben, J. A. in *The Ecology and Biology of Mammal-Like Reptiles* (eds Hotton, N., MacLean, P. D., Roth, J. J. & Roth, E. C.) 207–218 (Smithsonian Institution Press, 1986).
5. Castanet, J. *et al. J. Zool.* **263,** 31–39 (2004).
6. De Ricqlès, A. in *A Cold Look at the Warm Blooded Dinosaurs* (eds Thomas, R. D. K. & Olson, E. C.) 103–138 (Westview Press, 1980).
7. Padian, K., Horner, J. R. & De Ricqlès, A. *J. Vertebrate Paleontol.* **24,** 555–571 (2004).
8. Woodward, H. N., Horner, J. R. & Farlow, J. O. *J. Herpetol.* **45,** 339–342 (2011).
9. De Ricqlès, A., Padian, K. & Horner, J. R. *Annales de Paléontologie* **89,** 67–101 (2003).
10. Padian, K. & Lamm, E.-T. (eds) *Bone Histology of Fossil Tetrapods: Advancing Methods, Analysis, and Interpretation* (Univ. California Press, in the press).
11. Werning, S. *PLoS ONE* **7,** e33539 (2012).

ARTICLE

doi:10.1038/nature11229

Deep carbon export from a Southern Ocean iron-fertilized diatom bloom

Victor Smetacek[1,2]*, Christine Klaas[1]*, Volker H. Strass[1], Philipp Assmy[1,3], Marina Montresor[4], Boris Cisewski[1,5], Nicolas Savoye[6,7], Adrian Webb[8], Francesco d'Ovidio[9], Jesús M. Arrieta[10,11], Ulrich Bathmann[1,12], Richard Bellerby[13,14], Gry Mine Berg[15], Peter Croot[16,17], Santiago Gonzalez[10], Joachim Henjes[1,18], Gerhard J. Herndl[10,19], Linn J. Hoffmann[16], Harry Leach[20], Martin Losch[1], Matthew M. Mills[15], Craig Neill[13,21], Ilka Peeken[1,22], Rüdiger Röttgers[23], Oliver Sachs[1,24], Eberhard Sauter[1], Maike M. Schmidt[25], Jill Schwarz[1,26], Anja Terbrüggen[1] & Dieter Wolf-Gladrow[1]

Fertilization of the ocean by adding iron compounds has induced diatom-dominated phytoplankton blooms accompanied by considerable carbon dioxide drawdown in the ocean surface layer. However, because the fate of bloom biomass could not be adequately resolved in these experiments, the timescales of carbon sequestration from the atmosphere are uncertain. Here we report the results of a five-week experiment carried out in the closed core of a vertically coherent, mesoscale eddy of the Antarctic Circumpolar Current, during which we tracked sinking particles from the surface to the deep-sea floor. A large diatom bloom peaked in the fourth week after fertilization. This was followed by mass mortality of several diatom species that formed rapidly sinking, mucilaginous aggregates of entangled cells and chains. Taken together, multiple lines of evidence—although each with important uncertainties—lead us to conclude that at least half the bloom biomass sank far below a depth of 1,000 metres and that a substantial portion is likely to have reached the sea floor. Thus, iron-fertilized diatom blooms may sequester carbon for timescales of centuries in ocean bottom water and for longer in the sediments.

The Southern Ocean is regarded as a likely source and sink of atmospheric CO_2 over glacial–interglacial climate cycles, but the relative importance of physical and biological mechanisms driving CO_2 exchange are under debate[1,2]. The iron hypothesis[3], which is based on iron limitation of phytoplankton growth in extensive, nutrient-rich areas of today's oceans, is that the greater supply of iron-bearing dust to these regions during the dry glacials stimulated phytoplankton blooms that, by sinking from the surface to the deep ocean, sequestered climatically relevant amounts of carbon from exchange with the atmosphere. Twelve ocean iron fertilization (OIF) experiments carried out to test this hypothesis have provided unambiguous support for the first condition: that iron addition generates phytoplankton blooms in regions with high nutrient but low chlorophyll concentrations including the Southern Ocean[4,5]. The findings are consistent with satellite observations of natural phytoplankton blooms in these regions stimulated by dust input from continental[6] and volcanic[7] sources.

The timescales on which CO_2 taken up by phytoplankton is sequestered from the atmosphere depend on the depths at which organic matter sinking out of the surface layer is subsequently remineralized back to CO_2 by microbes and zooplankton. In the Southern Ocean, the portion of CO_2 retained within the 200-m-deep winter mixed layer would be in contact with the atmosphere within months, but carbon sinking to successively deeper layers, and finally the sediments, will be sequestered for decades to centuries or longer. Previous OIF experiments have not adequately demonstrated the fate and depth of sinking of bloom biomass[5], so it is uncertain whether mass, deep-sinking events comparable to those observed in the aftermath of natural blooms[8] also ensue from OIF blooms. Furthermore, palaeo-oceanographic proxies from the underlying sediments are ambiguous regarding productivity of the glacial Southern Ocean[1,2,9,10]. Hence, the second condition of the iron hypothesis, that OIF-generated biomass sinks to greater depths, has yet to be confirmed. The issue is currently receiving broad attention because OIF is one of the techniques listed in the geoengineering portfolio to mitigate the effects of climate change[11].

Monitoring the sinking flux from an experimental bloom requires vertical coherence between surface and deeper layers, a condition fulfilled by the closed cores of mesoscale eddies formed by meandering frontal jets of the Antarctic Circumpolar Current, which are prominent in satellite altimeter images as sea surface height anomalies[12]. An

[1]Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany. [2]National Institute of Oceanography, Dona Paula, Goa 403 004, India. [3]Norwegian Polar Institute, Fram Centre, Hjalmar Johansens Gate 14, 9296 Tromsø, Norway. [4]Ecology and Evolution of Plankton, Stazione Zoologica Anton Dohrn, Villa Comunale, 80121-Napoli, Italy. [5]Johann Heinrich von Thünen Institute, Institute of Sea Fisheries, Palmaille 9, 22767 Hamburg, Germany. [6]Department of Analytical and Environmental Chemistry, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [7]Univ. Bordeaux/CNRS, EPOC, UMR 5805, Station Marine d'Arcachon, 2 rue du Professeur Jolyet, F-33120 Arcachon, France. [8]Oceanography Department, University of Cape Town, Private Bag X3, Rondebosch, 7701 Cape Town, South Africa. [9]LOCEAN-IPSL, CNRS/UPMC/IRD/MNHN, 4 Place Jussieu, 75252 Paris Cedex 5, France. [10]Department of Biological Oceanography, Royal Netherlands Institute for Sea Research, PO Box 59, 1790 AB Den Burg, The Netherlands. [11]Department of Global Change Research, Instituto Mediterraneo de Estudios Avanzados, CSIC-UIB, Miquel Marques 21, 07190 Esporles, Mallorca, Spain. [12]Leibniz Institute for Baltic Sea Research Warnemünde, Seestraße 15, 18119 Rostock, Germany. [13]Bjerknes Centre for Climate Research, University of Bergen, Allegaten 55, N-5007 Bergen, Norway. [14]Norwegian Institute for Water Research, Thormøhlensgate 53 D, 5006 Bergen, Norway. [15]Department of Environmental Earth System Science, Stanford University, Stanford, California 94305, USA. [16]Helmholtz Centre for Ocean Research Kiel, Düsternbrooker Weg 20, 24105 Kiel, Germany. [17]Earth and Ocean Sciences, School of Natural Sciences, National University of Ireland, Galway, Quadrangle Building, University Road, Galway, Ireland. [18]Phytolutions GmbH, Campus Ring 1, 28759 Bremen, Germany. [19]Department of Marine Biology, University of Vienna, Althanstrasse 14, 1090 Vienna, Austria. [20]School of Environmental Sciences, University of Liverpool, Room 209 Nicholson Building, 4 Brownlow Street, Liverpool L69 3GP, UK. [21]Wealth from Oceans Flagship, Commonwealth Scientific and Industrial Research Organisation, Castray Esplanade, Hobart, Tasmania 7000, Australia. [22]MARUM – Center for Marine Environmental Sciences, University of Bremen, Leobener Strasse, D-28359 Bremen, Germany. [23]Institute for Coastal Research, Helmholtz-Zentrum Geesthacht, Center for Materials and Coastal Research, Max-Planck-Strasse 1, 21502 Geesthacht, Germany. [24]Eberhard & Partner AG, General Guisan Strasse 2, 5000 Arau, Switzerland. [25]Centre for Biomolecular Interactions Bremen, FB 2, University of Bremen, Postfach 33 04 40, 28359 Bremen, Germany. [26]School of Marine Science & Engineering, Plymouth University, Drake Circus, Plymouth PL4 8AA, UK.
*These authors contributed equally to this work.

added advantage offered by closed eddy cores is that processes occurring in the fertilized patch can be compared with natural processes in adjoining unfertilized waters of the same provenance.

The eddy and experiment

The European Iron Fertilization Experiment (EIFEX) was carried out from 11 February 2004 to 20 March 2004 during RV *Polarstern* cruise ANT XXI/3, in the clockwise-rotating core of an eddy formed by a meander of the Antarctic polar front (Fig. 1 and Supplementary Information). The eddy was mapped over a period of seven days shortly after fertilization of the patch with a grid of 80 stations along eight north–south transects 9 km apart. The rotating patch was encountered on two of the grid transects. Measurements of current speed and direction with the vessel-mounted acoustic Doppler current profiler and images of sea surface height anomalies revealed a closed, 60-km-diameter core clearly demarcated from the surrounding meander of the Antarctic polar front in all measured physical, chemical and biological properties (Fig. 1a, b and Supplementary Information).

Estimates of geostrophic shear and transports derived from temperature and salinity profiles (Supplementary Figs 1 and 2) indicate a coherent surface-to-bottom eddy circulation that was almost closed

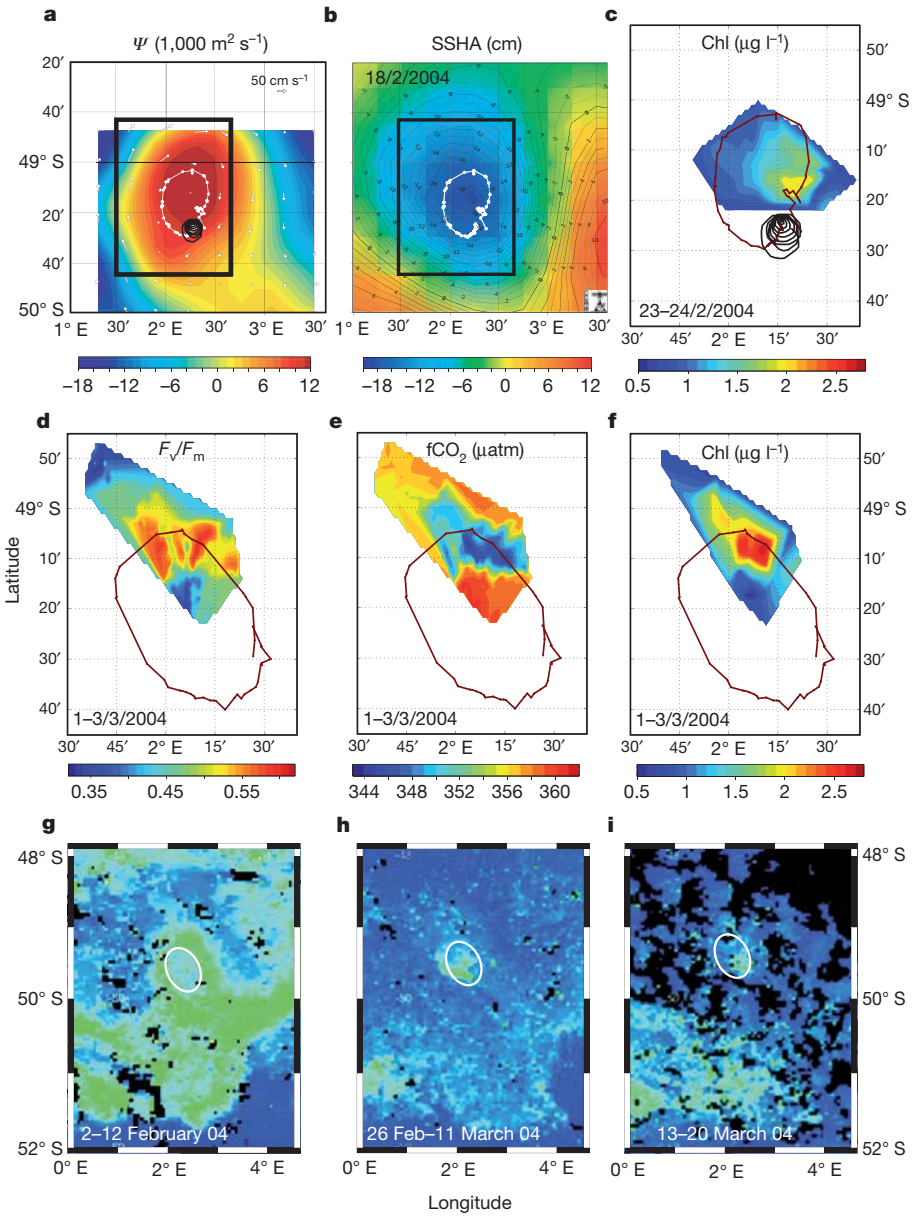


Figure 1 | Experimental eddy and the fertilized patch. **a**, The eddy core depicted with the stream function (Ψ; contours and colour scale) derived from currents measured using a vessel-mounted acoustic Doppler current profiler at a regular grid of stations between days 1 and 7. The black spiral is the ship's track (a Lagrangian circle) around the buoy drifting southwestward during fertilization. The white line is the superimposed track of the drifting buoy during its first rotation from days −1 to 11 (same as in **b** and **c**). **b**, Altimeter image of sea surface height anomaly (SSHA; contours and colour scale from CCAR, http://argo.colorado.edu/~realtime/gsfc_global-real-time_ssh/.) The rectangle in **a** and **b** is enlarged in **c**–**f**. **c**, Area and location of the patch on days 10 and 11 after fertilization, depicted on the basis of chlorophyll measurements. The yellow area is the hot spot. **d**–**f**, Location and area of the patch 17 days after fertilization, depicted in terms of photochemical efficiency (F_v/F_m; **d**), CO_2 fugacity (fCO_2; **e**) and chlorophyll concentration (**f**). The line is the track of the drifting buoy during its second rotation (days 13–21). The red area in **f** is the hot spot. **g**–**i**, Satellite-derived surface chlorophyll concentrations of the EIFEX eddy before fertilization (**g**), during the bloom peak (**h**) and in its demise phase (**i**). The eddy core is encircled in white; the EIFEX bloom is evident in **h** and **i** (green colour is >1 μg Chl l^{-1}). Note the natural bloom along the Antarctic polar front, which disappeared in this period. SeaWiFS images (**g**–**i**) courtesy of the NASA SeaWiFS Project and GeoEye.

and had little divergence. The vertical coherence of the eddy was also revealed by the congruent tracks of four neutrally buoyant floats positioned at respective depths of 200, 300, 500 and 1,000 m (Supplementary Fig. 3). A post-cruise Lagrangian analysis based on delayed-time altimetry[13] showed that, for the entire duration of the experiment, the compact core was only marginally eroded by lateral stirring, losing less than 10% of its content in total (Supplementary Information). This finding is consistent with diffusive heat budgets derived from the observed warming of the eddy's cold core[14]. Hence, the EIFEX eddy provided ideal conditions for monitoring the same water column from the surface to the sea floor over time.

The site of the pre-fertilization control station was marked with a drifting buoy around which a circular patch of 167 km^2 was fertilized with dissolved Fe(II) sulphate on 12–13 February (day 0) to yield a concentration of 1.5 µmol Fe m^{-3} in the 100-m-deep surface mixed layer, which is greater than background values by a factor of around five. A second fertilization on days 13 and 14 added an additional 0.34 µmol Fe m^{-3} to the 100-m-deep surface layer of the spreading patch. The patch was inadvertently placed off-centre but well within the closed eddy core and completed four rotations during the experiment. The area of the patch increased from 167 km^2 on day 0 to 447 km^2 on day 11 and to 798 km^2 on day 19 (Fig. 1c–f and Supplementary Information). 'In-stations' were taken in the least-diluted region of the patch: the 'hot spot' (Fig. 1c–f). 'Out-stations' were taken within the eddy core well away from the patch but in different locations relative to it, and hence did not represent ideal controls for quantifying processes within the patch.

Sampling frequency and depth coverage by discrete measurements are illustrated by the vertical distributions of chlorophyll and silicate concentrations (Fig. 2). Vertical profiles from *in situ* recording instruments indicated that the boundary of the mixed layer, defined by a sharp dip in temperature, salinity, fluorescence and transmission, was generally at a depth of 100 m (ref. 15; 97.6 ± 20.6 m). The element and biomass budgets presented here are based on inventories (in moles or grams per square metre) derived from the trapezoidal integration of six to eight discrete measurements from the 100-m-deep surface layer. For comparison with other studies, the stocks (inventories) from this layer are also presented where appropriate as depth-averaged concentrations (in millimoles or milligrams per cubic metre).

Processes inside and outside the patch

Enhanced phytoplankton growth stimulated by iron fertilization resulted in highly significant, linear increases in stocks of chlorophyll (Chl), particulate organic carbon (POC), nitrogen (PON), phosphate (POP) and biogenic silica (BSi) until day 24 (Figs 2a and 3). These stocks, depicted as depth-averaged concentrations in Fig. 3, declined thereafter, although at different rates. However, inventories of the corresponding dissolved nutrients, including dissolved organic nitrogen (DON), underwent a constant, linear decline until the end of the experiment, indicating that the respective uptake rates were maintained throughout (Fig. 3). Detailed, visual, quantitative examination of organisms and their remains across the entire size spectrum of the plankton revealed that population growth of many different species of large diatoms accounted for 97% of the Chl increase. The decline after day 24 was caused by mass death and formation of rapidly sinking aggregates by some diatom species, which was partly compensated by continued growth of other, heavily silicified species with high accumulation rates. The strikingly linear, instead of exponential, trends can be attributed to the effects of patch dilution with surrounding water because dilution rates (0.06–0.1 d^{-1}; Supplementary Information) and phytoplankton accumulation rates (0.03–0.11 d^{-1}) were similar.

The discrepancies in the budgets of particulate and dissolved pools of the various elements in the surface layer can only be explained by the sinking out of particles, as losses to dissolved organic pools can largely be ruled out: stocks of dissolved organic carbon (DOC) remained stable (44 ± 2 mmol C m^{-3}; Fig. 3l), those of DON halved (from 3.8 to 2.0 mmol N m^{-3}; Fig. 3f) and those of dissolved organic phosphate (not shown) were at the detection limit. The decline in DON was barely reflected in DOC because it was apparently associated with a relatively small, labile fraction with much lower C/N ratios than that of the large refractory DOC pool.

The post-fertilization eddy survey revealed that the patch was located in the region of the eddy core with the highest silicate and lowest chlorophyll concentrations (~0.7 mg Chl m^{-3}; Supplementary Information). Patchy, natural blooms, probably caused by local dust input along the Antarctic polar front[6], had occurred before our arrival adjacent to the patch (Fig. 1g), as indicated by lower nutrient and higher Chl stocks (up to 1.2 mg Chl m^{-3}) and higher particle loads in subsurface layers (Supplementary Information). These

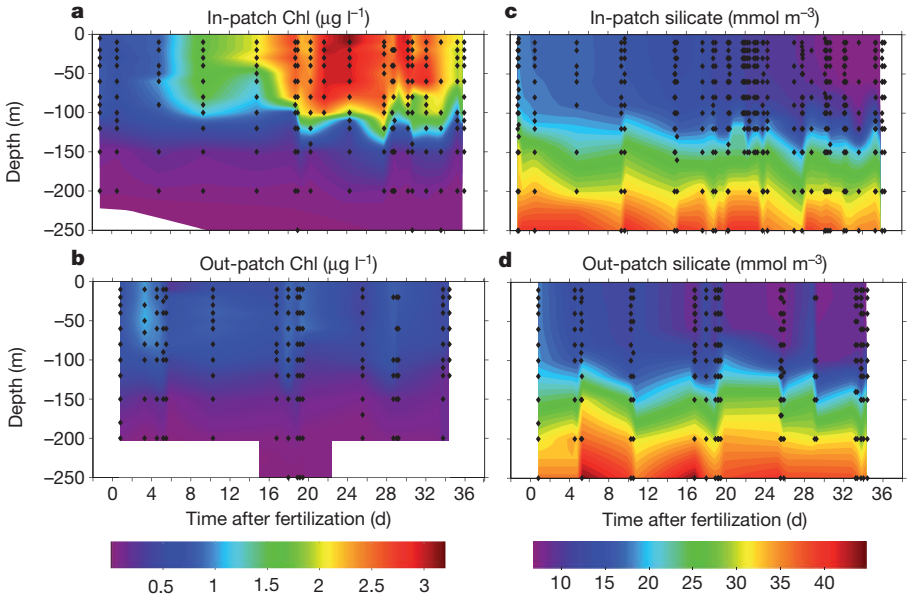


Figure 2 | Temporal evolution of chlorophyll and silicate concentrations. **a**, Chlorophyll concentrations reflect the growth, peak and demise phases of the bloom in the patch. **b**, By comparison, the Chl concentration outside the patch is low. The slightly higher out-patch values soon after fertilization are due to local patchiness in outside water and not to interim accumulation. **c**, **d**, The declining trend of silicate in outside water (**d**) is interrupted by local patchiness, whereas within the patch the trend is smooth (**c**). Note the variations in mixed-layer depth below 100 m. Black diamonds indicate depths of discrete samples.

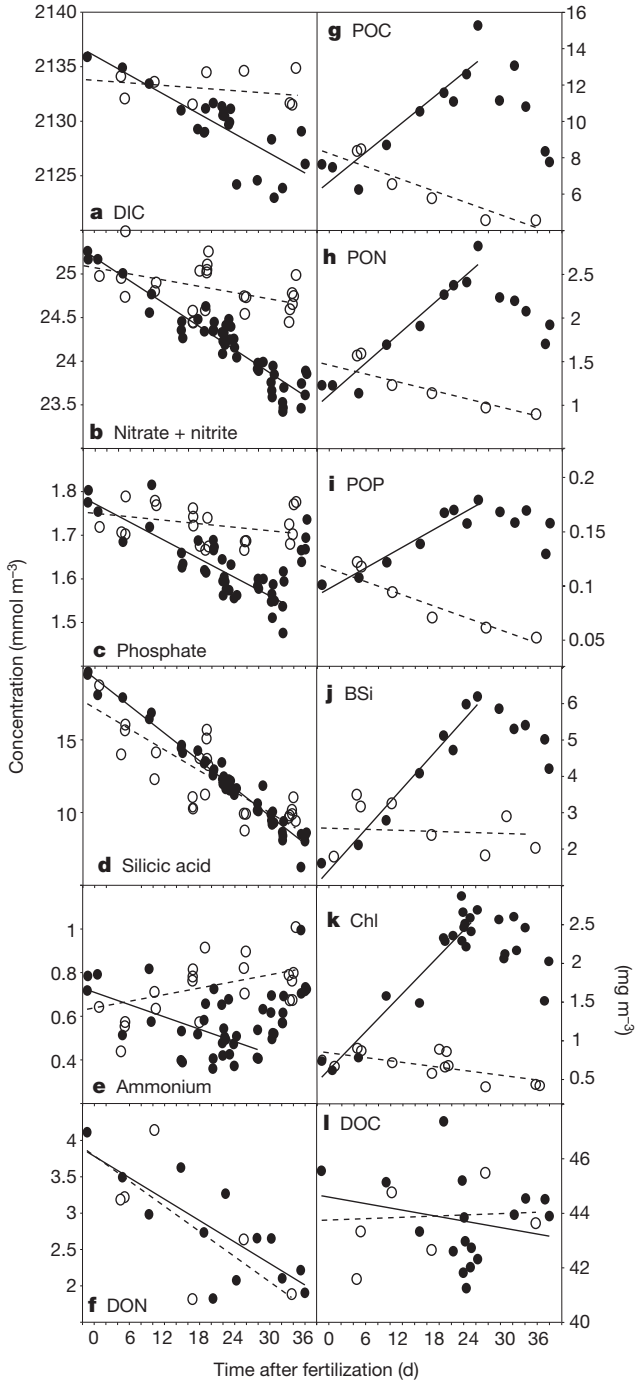


Figure 3 | Temporal evolution of dissolved and particulate elements. Values inside (filled circles) and outside (open circles) the fertilized patch are depth-integrated average concentrations for the upper 100 m of the water column. All concentrations are in millimoles per cubic metre except that for Chl (**k**), which is expressed in milligrams per cubic metre. Lines represent the temporal evolution inside (solid line) and outside (broken line) the fertilized patch used in elemental budget calculations (Supplementary Methods) determined by linear regression. Inside the patch, the r^2 values for the models are 0.64 (DIC; **a**), 0.88 (nitrate plus nitrite; **b**), 0.74 (phosphate; **c**), 0.97 (silicic acid; **d**), 0.33 (ammonium; **e**), 0.63 (DON; **f**), 0.84 (POC; **g**), 0.94 (PON; **h**), 0.93 (POP; **i**), 0.96 (BSi; **j**), 0.92 (Chl; **k**) and 0.05 (DOC; **l**). Outside the patch, the r^2 values are 0.06 (DIC), 0.24 (nitrate plus nitrite), 0.13 (phosphate), 0.71 (silicic acid), 0.24 (ammonium), 0.58 (DON), 0.84 (POC), 0.64 (PON), 0.85 (POP), 0.008 (BSi) and 0.005 (DOC). All regressions are significant ($P < 0.005$) with the exception of in-patch DOC ($P = 0.4$) and out-patch DIC ($P = 0.5$), nitrate plus nitrite ($P = 0.011$), phosphate ($P = 0.1$), ammonium ($P = 0.02$), DON ($P = 0.046$), PON ($P = 0.03$), BSi ($P = 0.8$) and DOC ($P = 0.8$).

natural blooms sank from the surface in the first week, as corroborated by barite profiles in subsurface layers[16], and, hence, greater particle stocks encountered at depth at some out-stations probably stemmed from them. Particle stocks (except those of BSi) in the upper 100-m-deep layer outside the patch continued declining by about 50% during the five weeks of the experiment (Fig. 3) as a result of steady sinking out of particles, as indicated by the discrepancy between the temporal evolutions of dissolved and particulate inventories of the respective elements. On a visit to the eddy core on day 50, we found that Chl concentrations had declined further, to 0.1 mg Chl m^{-3}.

Export from the iron-induced bloom

Vertical particle flux (export) induced by iron fertilization was estimated for the hot spot from losses of biogenic element inventories in the surface layer, including ^{234}Th; the increase in POC in the underlying deep water column; and by balancing rates of primary production and heterotrophic activity. These estimates represent the total export and include losses incurred by the surface layer in the absence of fertilization (background flux). Because some out-stations were affected by patchy natural blooms within the eddy core, they did not represent ideal controls. However, export rates from the fertilized patch, estimated from elemental budgets in the surface layer (0–100 m) and particularly POC increments between depths of 200 and 500 m until day 24, were remarkably similar to the corresponding values from outside water (Fig. 4), indicating little additional flux from the growing bloom. Export from the patch increased steeply after day 24, but outside loss rates remained constant. Hence, we assume that the background flux from the patch also remained constant until the end of the experiment but was overridden by the iron-induced flux event starting between days 24 and 28. We estimate the total background export from the patch by extrapolating the flux between days 0 and 24 to day 36, and subtract this amount from the total export to obtain the iron-induced export (Supplementary Methods).

Element losses

A conservative estimate for the export of biogenic elements from the 100-m-deep surface layer in the hot spot is the difference between the decline in dissolved inventories (nutrient uptake) and concomitant accumulation in particulate stocks (Supplementary Table 1). Uptake was calculated from the linear regressions depicted in Fig. 3a–f. Accumulation was estimated as the difference between final and initial stocks of BSi, POP, PON and POC. Initial stocks of particulate elements were taken from the intercept on day 0 of the linear regressions until day 24, and final stocks were the values measured on day 36 (Fig. 3g–j). The decline in the inventory of dissolved inorganic carbon (DIC), of 1.1 mol m^{-2}, underestimates the actual uptake by phytoplankton because of atmospheric CO_2 replenishment. Correcting for air–sea gas exchange adds 0.4 mol m^{-2} to the uptake and, hence, also to export. The background flux was obtained by extrapolating the losses estimated until day 24 (0.3 mol C m^{-2}) to the values on day 36. The iron-induced export, of 0.9 mol C m^{-2} (79 mmol C m^{-2} d^{-1} for the 12-day flux event), was obtained by subtracting the background flux from the total flux estimates (1.4 mol C m^{-2}) for the 36 days of the calculations (Supplementary Table 1).

Correcting budgets with the measured values presented above for mixed-layer deepening and diapycnal mixing using the diffusion coefficient of 3.3×10^{-4} m^2 s^{-1} estimated for the EIFEX eddy core[14] almost doubles the estimates of background export inside the hot spot (from 0.4 to 0.9 mol C m^{-2}, or 12 to 26 mmol C m^{-2} d^{-1}) and outside the fertilized patch (from 0.7 to 1.2 mol C m^{-2}, or 21 to 36 mmol C m^{-2} d^{-1}). Because nutrient input to the surface layer from below was relatively constant during the experiment, the correction for iron-induced export during the last 12 days was comparatively minor: from 0.9 to 1.1 mol C m^{-2}, or 79 to 98 mmol C m^{-2} d^{-1}. Furthermore, correcting the hot-spot budgets for the effects of

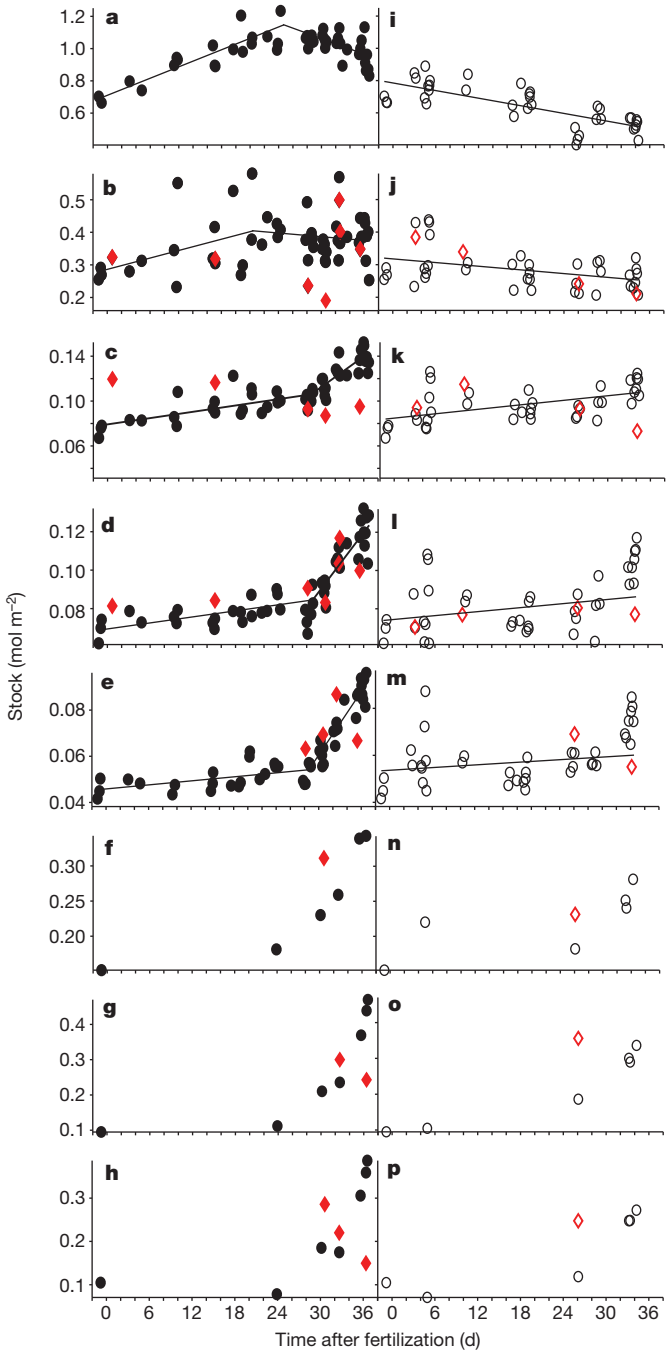


Figure 4 | Temporal evolution of particulate organic carbon stocks in successive depth layers. Stocks for the respective layers are derived from depth-integrated, vertical profiles of beam attenuation of a transmissometer calibrated using discrete POC measurements (black symbols). Filled and open symbols show data inside and outside the patch, respectively. Depth intervals of integrations are 0–100 m (**a**, **i**), 100–200 m (**b**, **j**), 200–300 m (**c**, **k**), 300–400 m (**d**, **l**), 400–500 m (**e**, **m**), 500–1,000 m (**f**, **n**), 1,000–2,000 m (**g**, **o**) and 2,000–3,000 m (**h**, **p**). Lines are derived from linear regression models. Variability in stocks and trends in the layers at 100–200 m (**b**, **j**) is due to intermittent shoaling and deepening of the particle-rich, surface mixed layer between 100 and 120 m, possibly as a result of the passage of internal waves. The high out-patch values on days 5 and 34 are not included in the regressions. The layer below 3,000 m is not included to avoid contamination by resuspended sediments in the nepheloid layer. Red diamonds show integrated stocks from measurements on discrete water samples. Variability in these values is due to low depth resolution, particularly below 500 m.

horizontal dilution (patch spreading) increases the total DIC uptake to 2.4 mol C m^{-2}, of which 0.6 mol C m^{-2} is exported laterally (Table 1). However, the increase of iron-induced export to 1.2 mol m^{-2} is again minor (Table 1). Dividing the total DIC uptake (2.4 mol C m^{-2}) by the total amount of iron added to the patch water column (0.18 mmol Fe m^{-2}) yields a C/Fe ratio of 13,000 ± 1,000 (s.e.m.). This ratio is conservative for reasons discussed in Supplementary Information.

Nitrate (nitrate plus nitrite) uptake until the bloom peak on day 24 (0.17 mol m^{-2}) accounted for 80% of PON production (0.21 mol N m^{-2}), resulting in negative values for background flux not indicated by the other elements (Supplementary Table 2). The decline in DON, through uptake by bacteria and excretion as ammonium to phytoplankton, more than compensates for the N deficit, but its origin is enigmatic. The same amount of DON, but much less DIC, nitrate and phosphate, were taken up outside the patch (Fig. 3 and Supplementary Table 3); hence DON contribution to export outside the patch is likely to have been similar to that inside the patch. The high variability in ammonium stocks (Fig. 3e) is consistent with rapid turnover within this pool. No clear trends were observed in the subsurface ammonium maxima inside and outside the patch (Supplementary Fig. 4), indicating minor additional accumulation of breakdown products from the flux event in the subsurface layer. The C/N ratio for iron-induced export (8.5) is higher than the POC/PON ratio of suspended elements (~5), a result that was also observed during the Southern Ocean Iron Experiment[4].

The steep increase in phosphate stocks on days 35 and 36 (Fig. 3c) is only partly explained by leaching from autolysed cytoplasm, owing to the well-established greater mobility of this element relative to carbon[17]. Hence, the negative value for exported P due to fertilization (Table 1) is difficult to explain, because unlike for N, the source of the additional phosphate during the last two days is unknown. Approximately 65% of the silicate taken up was exported during the 36 days, of which half can be attributed to iron-induced export at a C/Si ratio of 3 and the other half to background flux at a C/Si ratio of 0.8. Outside the patch, silicate uptake was slightly lower than inside but all of it was exported in the same period at a C/Si ratio of 0.9, which we attribute to the activity of diatom species that selectively sink silica.

POC export rates from the hot spot estimated from ^{234}Th increased steeply from background values <40 mmol C m^{-2} d^{-1} to 125 mmol C m^{-2} d^{-1} during days 28–32, but declined thereafter presumably owing to uncertainties associated with short-term sampling by this method (Supplementary Information). Nevertheless, the two peak values during the flux event are the highest recorded so far in the Southern Ocean.

Transmissometer profiles

Beam attenuation of the profiling transmissometer was highly correlated with discrete POC measurements across the entire range of concentrations encountered ($r^2 = 0.934$ and $P < 0.001$, where r is the correlation coefficient and P is the observed significance level; Supplementary Fig. 5). Because of the high-resolution vertical coverage of the water column, the depth-integrated transmissometer profiles provide a record of POC accumulation and depletion over depth and time from which export can be estimated[18].

Integrated POC stocks in the upper 3,000 m of the water column of the patch increased over 36 days by 1.3 ± 0.2 mol C m^{-2} (s.d.; Fig. 5), implying an accumulation rate of 38 mmol C m^{-2} d^{-1}. The flux event after day 24 is signalled by steeply increasing POC stocks in the water column below 200 m (Fig. 4). These stocks reached 0.8 ± 0.1 mol C m^{-2} (s.d.) above background levels on day 36, of which 0.7 ± 0.1 mol C m^{-2} (s.d.) was below 500 m (Fig. 5). The increase in deep POC stocks is reasonably close to the corrected estimate of iron-induced export from the surface layer budget (1.2 ± 0.4 mol C m^{-2} (s.d.)), given that some POC had already reached the deep-sea floor, as indicated by fresh diatom cells and labile pigments found close to the bottom (Supplementary Fig. 7). Hence, losses of iron-induced sinking flux due to ongoing respiration were apparently minor. Profiles of biogenic

Table 1 | Total and iron-induced export from the hot spot of the fertilized patch

	Days	Si	P	$NO_2 + NO_3$	Total N	C
Decrease in stocks of dissolved elements	0–36	1.14 ± 0.03	0.007 ± 0.003	0.160 ± 0.008	0.33 ± 0.08	1.1 ± 0.2
Input due to air–sea gas exchange	0–36	—	—	—	—	0.4
Dissolved-element input from vertical mixing (diapycnal mixing and deepening of the mixing layer)	0–36	0.22 ± 0.01	0.011 ± 0.0008	0.059 ± 0.004	0.10 ± 0.01	0.6 ± 0.2
Dissolved-element input from horizontal mixing (dilution effect)	0–36	~0	0.010 ± 0.001	0.061 ± 0.005	0.08 ± 0.03	0.27 ± 0.06
Total uptake (decrease in dissolved stocks plus gas exchange plus vertical mixing plus horizontal mixing)	0–36	1.36 ± 0.03	0.028 ± 0.003	0.28 ± 0.01	0.50 ± 0.09	2.4 ± 0.2
Difference between final and initial particulate-matter standing stocks	0–36	0.28 ± 0.02	0.0062 ± 0.0005	0.081 ± 0.009	0.081 ± 0.009	0.11 ± 0.07
Particulate matter loss by horizontal mixing (dilution effect)	0–36	0.19 ± 0.02	0.0089 ± 0.0004	0.110 ± 0.006	0.110 ± 0.006	0.58 ± 0.04
Vertical export (total uptake minus difference in particulate stocks minus particulate loss by horizontal mixing)	0–36	0.89 ± 0.04	0.013 ± 0.003	0.09 ± 0.02	0.31 ± 0.09	1.7 ± 0.2
Background (vertical) export*	0–36	0.50 ± 0.07	0.025 ± 0.003	−0.06 ± 0.03	0.18 ± 0.06	0.5 ± 0.3
Vertical export due to fertilization (vertical export minus background export)	24–36	0.40 ± 0.08	−0.012 ± 0.005	ND	0.1 ± 0.1	1.2 ± 0.4

Budgets were calculated for the surface layer (0–100 m) starting immediately after fertilization (day 0) and lasting until the last station where nutrients were measured inside the fertilized patch (day 36) (see Supplementary Methods for details). Total N includes $NO_2 + NO_3$, DON and ammonium. All values are in moles per square metre. Uncertainties (s.e.m.) were estimated by propagation of standard errors based on linear uptake models (Fig. 3) and measurement uncertainties. ND, not determined.
* Vertical export between days 0 and 24 (0.4 mol C m^{-2}; Supplementary Table 3) extrapolated to days 0–36 (Supplementary Methods).

barite[16] indicate that only ~11% of POC exported during the flux event was remineralized between depths of 200 and 1,000 m. By contrast, the background export until day 24 was largely respired above 500 m, because the POC increase below 200 m (Fig. 4) amounted to 0.04 mol C m^{-2}, which is <10% of the concomitant loss from the surface layer calculated from corrected element budgets. Identical rates of POC increase in subsurface layers outside the patch also indicate that background export was remineralized above 500 m.

We attribute the comparatively high POC stocks in outside waters (0.5 ± 0.1 mol C m^{-2} (s.d.) above background) between 200 and 1,000 m on day 5 and between 300 and 3,000 m on days 33 and 34 (Fig. 4) to slower sinking particle flux from the patchy natural blooms mentioned above (Supplementary Information). Local patchiness in out-stations is indicated by greater scatter in values from successive profiles taken during the same station than in in-stations (Figs 4 and 5).

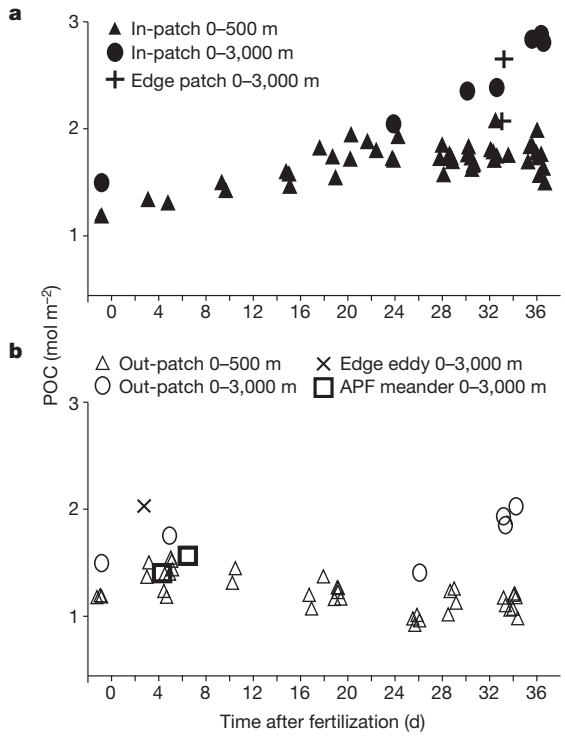


Figure 5 | Depth-integrated particulate organic carbon stocks. Stocks for the 0–500-m (triangles) and 0–3,000-m (circles) water columns are derived from vertical transmissometer profiles as in Fig. 4. Filled and open symbols show data inside and outside the patch, respectively. All profiles measured down to the sea floor during the study are depicted. Because the depth of the flux event was not anticipated and deep casts are time consuming, only six profiles to the sea floor were measured before the flux event: one before fertilization, one in the hot spot, two outside the patch (of which one was inside the core) and two in the meander of the Antarctic polar front (APF).

The steep increase in POC stocks below 200 m under the patch after day 28 (Fig. 4) can be attributed to POC in the mucilaginous matrix of diatom aggregates in the >1-mm size range, which appeared as spikes in the transmissometer profiles. The occurrence of large aggregates reflected in spikiness of the profiles was notably higher under the hot spot than outside it (Supplementary Fig. 6). Sinking rates of >500 m d^{-1} and aggregates in the centimetre size range are required to account for the similar slopes of increasing POC at all depths down to the sea floor after day 28, just four days after the enhanced appearance of spikes (aggregate formation) at the pycnocline (Supplementary Fig. 6a) and visual observation of mass mortality in a major, spiny diatom species, *Chaetoceros dichaeta*. By contrast, the smaller aggregates from the less dense natural blooms sank more slowly than those from the patch. Coagulation models[19,20] of aggregate formation confirm the relationships between sinking rate and bloom density and, respectively, cell size (including spine length). The latter depends on the species composition of the bloom, which thus has a decisive role in the long-term fate of its biomass.

Organism stocks and rates

Biomass estimates from organism counts were highly correlated with the respective bulk measurements: the ratio of POC to phytoplankton carbon was 1.4 mol mol^{-1} ($r^2 = 0.54$, $P < 0.0001$), that of phytoplankton carbon to Chl was 27 mg mg^{-1} ($r^2 = 0.87$, $P < 0.0001$) and that of BSi to total diatom carbon was 0.9 mol mol^{-1} ($r^2 = 0.76$, $P < 0.0001$). This indicated that accumulation of particulate organic detritus was minor, that POC increments had a high cellular chlorophyll content as a result of newly accumulated biomass (POC/Chl = 32 mg mg^{-1}, $r^2 = 0.394$, $P < 0.0001$) and that new biomass was dominated by diatoms. Rates of primary production doubled following fertilization and stabilized at 0.13 mol C m^{-2} d^{-1} (1.5 ± 0.1 g C m^{-2} d^{-1}) from days 9 to 35 but declined to 0.08 mol C m^{-2} d^{-1} on day 36. The total C accumulation and export rates estimated from element budgets are easily accommodated in the organic carbon produced during the bloom (4.2 mol C m^{-2}, or 51 g C m^{-2}). The remainder could be attributed to recycling in the surface layer by bacteria, microzooplankton and copepods, whereby grazing pressure on the diatoms was lower than on other protists. Bacterial stocks and production rates declined by ~30% in the demise phase of the bloom, which supports the high transfer efficiency of POC to depth. Thus, the budgets of biological rates are remarkably consistent with other budgets.

In outside water, primary production amounted to 1.36 mol C m^{-2} (16 g C m^{-2}) over 34 days, which is too low to accommodate the estimated sinking losses of 1.2 mol C m^{-2} (14.4 g C m^{-2}; Supplementary Table 3) because bacterial remineralization rates and copepod grazing pressure were in the same range outside the patch as inside. The

discrepancy between measured production and estimated loss rates is partly due to three out-stations placed where there had been previous natural blooms with comparatively low surface DIC inventories (Fig. 3a) but where much of the corresponding POC stocks had already sunk to subsurface layers. Furthermore, the vertical diffusion coefficient applied[14] seems to lead to an overestimation of export. Applying another, twofold-higher, diffusion coefficient for diapycnal mixing derived from microstructure profiles during EIFEX[15] results in twofold-higher export values that are supported even less by direct observations of the plankton community and by POC profiles.

Conclusions

The peak chlorophyll stock of 286 mg m^{-2} is the highest recorded in an OIF experiment so far[5] and demonstrates that, contrary to the current view[21], a massive bloom can develop in a mixed layer as deep as 100 m. The EIFEX results provide support for the second condition of the iron hypothesis[4], that mass sinking of aggregated cells and chains in the demise phase of diatom blooms also occurs in the open Southern Ocean, both in natural[22,23] and in artificially fertilized blooms. Given the large sizes, high sinking rates and low respiratory losses of aggregates from the iron-induced bloom, much of the biomass is likely to have been deposited on the sea floor as a fluff layer[24] with carbon sequestration times of many centuries and longer. Larger-scale, longer-term OIF experiments will be required to reduce the effects of horizontal dilution and to explore further the potential of this technique for hypothesis testing in the fields of ecology, biogeochemistry and climate.

METHODS SUMMARY

The eddy was selected on the basis of satellite altimetry and surface chlorophyll distribution. Fertilization was carried out by releasing 7 t of commercial Fe(II) sulphate dissolved in 54 m^3 of acidified (HCl) sea water into the ship's propeller wash while spiralling out from a drifting buoy at 0.9-km radial intervals. By day 14, the initial 167-km^2 patch had spread and an area of 740 km^2 was again fertilized with 7 t of Fe(II) sulphate, this time along east–west transects 3 km apart, from north to south in the direction of the moving patch (Fig. 1c).

The patch was located using the drifting buoy, and the photochemical efficiency (F_v/F_m) was measured continuously with a fast-repetition-rate fluorometer. As in previous experiments, F_v/F_m was significantly higher in iron-fertilized water. Within a week, the bloom had accumulated sufficient biomass that additional tracers (chlorophyll concentration and continuous measurements of fCO_2) could be used to locate the part of the patch least affected by dilution with outside water, that is, that with the highest chlorophyll concentration and, in the last week, the lowest fCO_2 value (Fig. 1c–f). All in-stations were placed inside this hot spot and care was taken to locate it with small-scale surveys before sampling and to keep the ship within it during sampling at each station, which generally lasted about 8 h. Some in-stations were within the patch but were subsequently shown to have missed the hot spot and have therefore been excluded. For logistical reasons, the out-stations were taken in different locations of the core relative to the direction of the moving patch, that is, ahead, behind or diagonally opposite it.

Standard oceanographic methods and instruments[15] were used to collect samples and measure the properties of the water column. See Supplementary Information for details.

Received 15 November 2011; accepted 3 May 2012.

1. Sigman, D. M. Hain, M. P. & Haug, G. H. The polar ocean and glacial cycles in atmospheric CO_2 concentration. *Nature* **466,** 47–55 (2010).
2. Anderson, R. F. *et al.* Wind-driven upwelling in the Southern Ocean and the deglacial rise in atmospheric CO_2. *Science* **323,** 1443–1448 (2009).
3. Martin, J. H. Glacial-interglacial CO_2 changes: the iron hypothesis. *Paleoceanography* **5,** 1–13 (1990).
4. Coale, K. H. *et al.* Southern Ocean iron enrichment experiment: carbon cycling in high- and low-Si waters. *Science* **304,** 408–414 (2004).
5. Boyd, P. *et al.* Mesoscale iron-enrichment experiments 1993–2005: synthesis and future directions. *Science* **315,** 612–617 (2007).
6. Cassar, N. *et al.* The Southern Ocean biological response to aeolian iron deposition. *Science* **317,** 1067–1070 (2007).
7. Hamme, R. C. *et al.* Volcanic ash fuels anomalous plankton bloom in subarctic northeast Pacific. *Geophys. Res. Lett.* **37,** L19604 (2010).
8. Lampitt, R. S. *et al.* Material supply to the abyssal seafloor in the Northeast Atlantic. *Prog. Oceanogr.* **50,** 27–63 (2001).
9. Abelmann, A., Gersonde, R., Cortese, G., Kuhn, G. & Smetacek, V. Extensive phytoplankton blooms in the Atlantic sector of the glacial Southern Ocean. *Paleoceanography* **21,** PA1013 (2006).
10. Kohfeld, K. E., Le Quéré, C., Harrison, S. P. & Anderson, R. F. Role of marine biology in glacial-interglacial CO_2 cycles. *Science* **308,** 74–78 (2005).
11. The Royal Society. *Geoengineering the Climate: Science, Governance and Uncertainty*. RS policy document 10/09 (The Royal Society, 2009).
12. Chelton, D. B., Schlax, M. G., Samelson, R. M. & de Szoeke, R. A. Global observations of large oceanic eddies. *Geophys. Res. Lett.* **34,** L15606 (2007).
13. d'Ovidio, F., Isern-Fontanet, J., Lopez, C., Hernandez-Garcia, E. & Garcia-Ladona, E. Comparison between Eulerian diagnostics and finite-size Lyapunov exponents computed from altimetry in the Algerian basin. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **56,** 15–31 (2009).
14. Hibbert, A., Leach, H., Strass, V. & Cisewski, B. Mixing in cyclonic eddies in the Antarctic Circumpolar Current. *J. Mar. Res.* **67,** 1–23 (2009).
15. Cisewski, B., Strass, V. H., Losch, M. & Prandke, H. Mixed layer analysis of a mesoscale eddy in the Antarctic Polar Front Zone. *J. Geophys. Res.* **113,** C05017 (2008).
16. Jacquet, S. H. M., Savoye, N., Dehairs, F., Strass, V. H. & Cardinal, D. D. Mesopelagic carbon remineralization during the European Iron Fertilization Experiment. *Glob. Biogeochem. Cycles* **22,** GB1023 (2008).
17. Paytan, A. & McLaughlin, K. The oceanic phosphorus cycle. *Chem. Rev.* **107,** 563–576 (2007).
18. Bishop, J. K. B., Wood, T. J., Davis, R. E. & Sherman, J. T. Robotic observations of enhanced carbon biomass and export at 55° S during SOFeX. *Science* **304,** 417–420 (2004).
19. Jackson, G. A. A model of the formation of marine algal flocs by physical coagulation processes. *Deep-Sea Res.* **37,** 1197–1211 (1990).
20. Riebesell, U. & Wolf-Gladrow, D. A. The relationship between physical aggregation of phytoplankton and particle flux: a numerical model. *Deep-Sea Res. A* **39,** 1085–1102 (1992).
21. de Baar, H. J. W. *et al.* Synthesis of iron fertilization experiments: from the iron age in the age of enlightenment. *J. Geophys. Res.* **110,** C09S16 (2005).
22. Blain, S. *et al.* Effect of natural iron fertilization on carbon sequestration in the Southern Ocean. *Nature* **446,** 1070–1074 (2007).
23. Pollard, R. *et al.* Southern Ocean deep-water carbon export enhanced by natural iron fertilization. *Nature* **457,** 577–580 (2009).
24. Beaulieu, S. E. in *Oceanography and Marine Biology. An Annual Review* (eds Gibson, R. N., Barnes, M. & Atkinson, R. J.) 171–232 (Taylor & Francis, 2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Balt, K. Loquay, S. Mkatshwa, H. Prandke, H. Rohr, M. Thomas and I. Vöge for help on board. We are also grateful to U. Struck for POC and PON analyses. The altimeter products were produced by Ssalto/Duacs and distributed by Aviso with support from Cnes. We thank the captain and crew of RV *Polarstern* (cruise ANT XXI/3) for support throughout the cruise.

Author Contributions V.S. and C.K. wrote the manuscript. V.S. directed the experiment and C.K. carried out the budget calculations. V.H.S., P.A., M.M. and D.W.-G. contributed to the preparation of the manuscript. V.H.S., B.C., H.L. and M.L. contributed physical data on mixed-layer depth dynamics, eddy coherence, patch movement and transmissometer data. N.S. provided thorium data. A.W. provided nutrient data. P.A. and J.H. provided phytoplankton and BSi data. F.D. carried out the Lagrangian analysis based on delayed-time altimetry. J.M.A. and G.J.H. provided bacterial data. C.N. and R.B. provided inorganic carbon data. G.M.B., C.K. and M.M.M. provided POC and PON data. P.C. provided the iron data. S.G. and A.T. provided DOM data. I.P. and L.J.H. performed the ^{14}C primary production measurements and provided high-pressure liquid chromatography data. R.R. provided data on photochemical efficiency (F_v/F_m). C.K., M.M.S. and A.T. provided Chl data. U.B., E.S., O.S. and J.S. provided data on the eddy core from a subsequent cruise and satellite Chl images.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to V.S. (victor.smetacek@awi.de) or C.K. (christine.klaas@awi.de).

ARTICLE

doi:10.1038/nature11251

Non-invasive prenatal measurement of the fetal genome

H. Christina Fan[1]†*, Wei Gu[1]*, Jianbin Wang[1], Yair J. Blumenfeld[2], Yasser Y. El-Sayed[2] & Stephen R. Quake[1,3,4]

The vast majority of prenatal genetic testing requires invasive sampling. However, this poses a risk to the fetus, so one must make a decision that weighs the desire for genetic information against the risk of an adverse outcome due to hazards of the testing process. These issues are not required to be coupled, and it would be desirable to discover genetic information about the fetus without incurring a health risk. Here we demonstrate that it is possible to non-invasively sequence the entire prenatal genome. Our results show that molecular counting of parental haplotypes in maternal plasma by shotgun sequencing of maternal plasma DNA allows the inherited fetal genome to be deciphered non-invasively. We also applied the counting principle directly to each allele in the fetal exome by performing exome capture on maternal plasma DNA before shotgun sequencing. This approach enables non-invasive exome screening of clinically relevant and deleterious alleles that were paternally inherited or had arisen as *de novo* germline mutations, and complements the haplotype counting approach to provide a comprehensive view of the fetal genome. Non-invasive determination of the fetal genome may ultimately facilitate the diagnosis of all inherited and *de novo* genetic disease.

Our work is based on the phenomenon of circulating cell-free DNA, whose existence and role in pregnancy was first investigated in 1948[1]. A portion of the cell-free DNA in a pregnant woman's blood is derived from the fetus[2], and this fact has enabled the development of a number of non-invasive prenatal diagnostic techniques[3]. A prominent example is the non-invasive detection of Down syndrome and other aneuploidies, which was first demonstrated by our group[4], validated by clinical trials[5–10], and is now available in the clinic. We describe here how the chromosome counting principle we invented for aneuploidy detection can be applied to non-invasive fetal genome analysis by directly counting haplotypes and even individual alleles. Others have studied the relationship between maternal and fetal cell-free DNA[11], but their approach required invasively sampled fetal material, did not determine the fetal genome, and also needed knowledge of paternal genetic data.

Measuring the fetal genome by counting parental haplotypes

Maternal plasma DNA is a mixture of maternal and fetal DNA; the fraction of fetal DNA ranges from a few percent or lower early in pregnancy to as high as ~50%[2,7], and generally increases with gestational age. Because the fetal genome is a combination of the four parental chromosomes, or haplotypes, as a result of random assortment and recombination during meiosis, three haplotypes exist in maternal plasma per genomic region: the maternal haplotype that is transmitted to the fetus, the maternal haplotype that is not transmitted, and the paternal haplotype that is transmitted. If the relative copy number of the untransmitted maternal haplotype is $1-\varepsilon$, where ε is the fetal DNA fraction, then the relative copy number of the transmitted maternal haplotype is 1, and the relative copy numbers of the transmitted and untransmitted paternal haplotypes are ε and 0, respectively (Fig. 1). Therefore, within each pair of parental haplotypes, the transmitted haplotype is over-represented relative to the untransmitted one. By measuring the relative amount of parental haplotypes through counting the number of alleles specific to each parental haplotype (referred to as 'markers'), one can deduce the inheritance of each parental haplotype and hence build the full inherited fetal genome.

Strictly speaking, the markers that define each maternal haplotype are the alleles that are present in one maternal haplotype but not in the other maternal haplotype and the two paternal haplotypes. However, because it is rare that two unrelated persons share the same long-range haplotype, that is, a haplotype much longer than the usual length of haplotype blocks observed in the population (~100 kilobases (kb)), the presence of alleles contributed by the transmitted paternal haplotype at these loci would not interfere with the measurement of representation of maternal haplotypes as long as the haplotype being considered is sufficiently long (>1 megabase (Mb)). Thus all the maternal heterozygous loci can be used to define the two maternal haplotypes (Fig. 1). This enables the measurement of relative representation of the two maternal haplotypes without the knowledge of paternal haplotypes. The relative representation of the two maternal haplotypes is the difference in the counts of markers specific to each haplotype. Even if the over-representation of the transmitted maternal haplotype is small, the over-represented haplotype can be identified provided that the counting depth exceeds the counting noise, which is governed by Poisson statistics. Supplementary Table 1 and Supplementary Fig. 1 provide estimations of counting requirement as a function of confidence of measurement and fetal DNA percentage in the clinically observed range. Because the number of markers that define each parental haplotype increases with haplotype length, the longer the phased haplotypes, the lower the average number of sampling per individual marker is required for confident determination of the over-represented parental haplotypes.

If paternal haplotypes are known, it is straightforward to determine the inherited paternal haplotypes by comparing the sum of count of

[1]Department of Bioengineering, Stanford University, Clark Center Rm E300, 318 Campus Drive, Stanford, California 94305, USA. [2]Division of Maternal-Fetal Medicine, Department of Obstetrics & Gynecology, Stanford University School of Medicine, 300 Pasteur Drive, Room HH333, Stanford, California 94305, USA. [3]Department of Applied Physics, Stanford University, Clark Center Room E300, 318 Campus Drive, Stanford, California 94305, USA. [4]Howard Hughes Medical Institute, Stanford University, Clark Center Room E300, 318 Campus Drive, Stanford, California 94305, USA.
†Current address: ImmuMetrix LLC, 552 Del Rey Avenue, Sunnyvale, California 94085, USA.
*These authors contributed equally to this work.

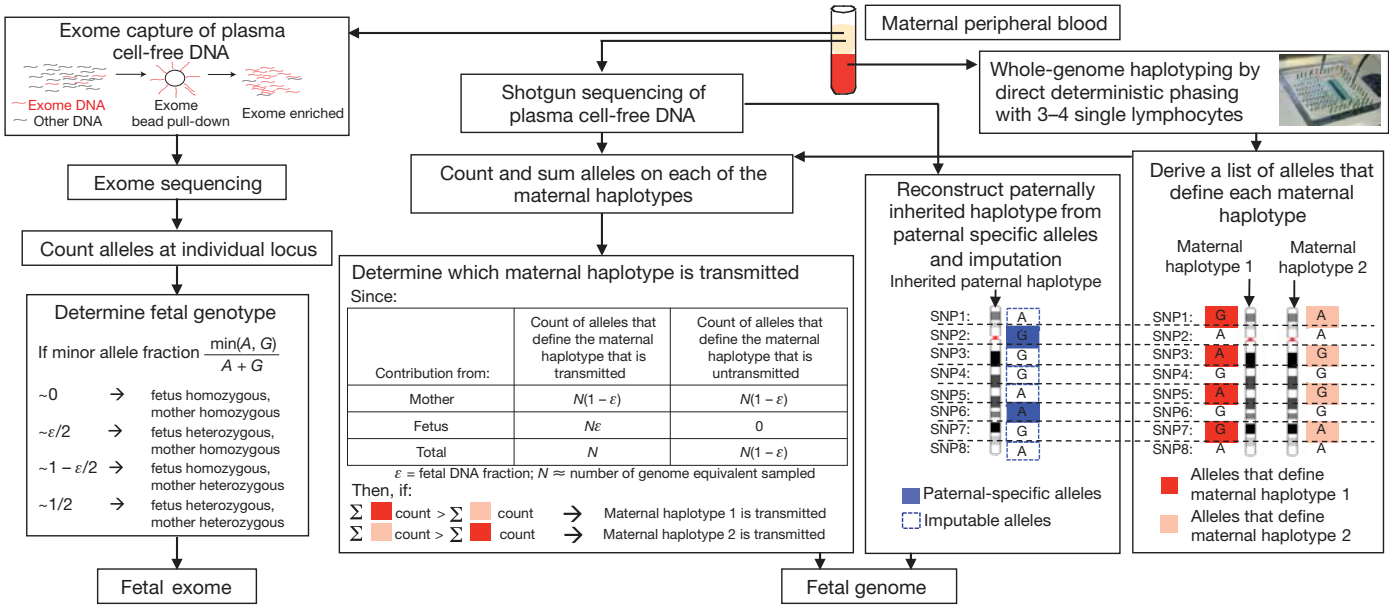


Figure 1 | Molecular counting strategies for measuring the fetal genome non-invasively from maternal blood only. Genome-wide, chromosome length haplotypes of the mother are obtained using direct deterministic phasing. The inheritance of maternal haplotypes is revealed by sequencing maternal plasma DNA and summing the count of the alleles specific to each haplotype at heterozygous loci and determining the relative representation of the two alleles. The inherited paternal haplotypes are defined by the paternal-specific alleles (that is, those that are different from the maternal ones at positions where the mother is homozygous). The allelic identity at loci linked to the paternal-specific alleles on the paternal haplotype can be imputed. Alternatively, molecular counting can be applied directly to count alleles at individual loci to determine fetal genotypes via targeted deep sequencing, such as exome-enriched sequencing of maternal plasma DNA. For illustrative purpose, each locus is biallelic and carries the 'A' or 'G' alleles.

alleles specific to each paternal haplotype (Supplementary Fig. 2), thereby revealing the entire inherited fetal genome. Supplementary Fig. 3 and the accompanying Supplementary Information show how this could be achieved using sequencing data of a synthetic mixture of DNA from a mother and daughter within a fully phased family trio[12]. However, it is not always possible to obtain paternal information; the incidence of non-paternity is estimated to be between 3% and 10%[13,14], making this a particularly delicate issue. In the absence of paternal information, the paternally inherited haplotypes can be reconstructed via linkage to observed non-maternal (that is, paternal-specific) alleles (Fig. 1).

We verified this approach on samples collected from two pregnancies. Pregnant woman P1 carried a female fetus with normal karyotype, whereas pregnant woman P2 is an individual with a ~2.85 Mb heterozygous deletion on chromosome 22 that is associated with DiGeorge syndrome. To obtain phased maternal chromosomes, we performed direct deterministic phasing (DDP)[15] on three or four maternal metaphase cells obtained by culturing maternal whole blood (Supplementary Table 2 and Supplementary Fig. 4). DDP involves microfluidic separation and amplification of individual metaphase chromosomes from single cells followed by genome-wide genotyping analysis of amplified materials, and enables each chromosome in the genome to be phased along its full length. Genomic DNA of cord blood collected at delivery was also genotyped to serve as the true reference for fetal genotypes. The true inheritance of maternal haplotypes was determined by aligning the homozygous SNPs of the fetus by cord blood genotyping against the two maternal haplotypes defined by the phased maternal heterozygous SNPs (Fig. 2). The analysis here concerns the approximately 1 million positions across the genome present on Omni1-Quad genotyping array. Phase information of the remaining genomic positions, particularly those that carry rare variants of clinical importance, can be obtained by broader array coverage or direct sequencing of amplified chromosome materials, as demonstrated previously[15].

Maternal cell-free DNA samples were shotgun-sequenced on the Illumina platform to a depth of ~52.7× (151 gigabases (Gb)), ~20.8× (59.7 Gb) and ~1.3× (3.7 Gb) haploid genome coverage for P1T1 (P1, 1st trimester), P1T2 (P2, 2nd trimester) and P2T3 (P3, 3rd trimester), respectively (Supplementary Table 2). To determine fetal inheritance of maternal haplotypes, we divided each chromosome into bins of 2.5–3.5 Mb for autosomal chromosomes and 5–7.5 Mb for chromosome X (Supplementary Table 2), with sliding steps of 100 kb, and compared the counts of alleles specific to each of the two haplotypes. Bin sizes were chosen according to the estimated sampling requirement (Supplementary Table 1) based on the sequencing depth, density of markers and fetal DNA fraction, which was estimated to be ~6%, ~16% and ~30% for P1T1, P1T2 and P2T3 by comparing relative representation of maternal haplotypes, respectively. The lower SNP array density on chromosome X required larger bin sizes for that chromosome. The over-represented maternal haplotype over the entire genome was apparent and corresponded to the maternal haplotype transmitted to the fetus (Fig. 2). Taking into account the uncertainty surrounding regions of crossovers (median ~350–450 kb per crossover, Supplementary Fig. 5), maternal inheritance of at least 99.2% of the SNPs could be deduced with at least 99.8% accuracy for all samples. Less sequencing depth also allowed the inherited maternal haplotypes to be deduced (Supplementary Fig. 6) with lower resolution of crossovers due to larger bin sizes (Supplementary Fig. 5).

The paternally inherited haplotypes were reconstructed by detection of paternal-specific alleles, followed by imputation at linked positions. We used the haplotypes of normal population documented by the 1000 Genome Project[16] as reference haplotypes for imputation. Imputation accuracy is dependent on the density of markers, and the number of identified non-maternal alleles is dependent on sequencing depth and fetal DNA fraction. At the final sequencing depth (~52.7×, ~20.8× and ~10.7× haploid genome coverage for P1T1, P1T2 and P2T3, respectively), we detected ~66–70% of the paternal-specific alleles at least once (Supplementary Table 2 and Supplementary Fig. 7). Approximately 3.4–5.6% of the non-maternal alleles were sequencing noise. Using the non-maternal markers, we deduced ~70% of the paternally inherited haplotypes with ~94–97% accuracy via imputation (Fig. 3). The loci that could not be confidently imputed reside in regions where paternal-specific alleles were not detected, in regions

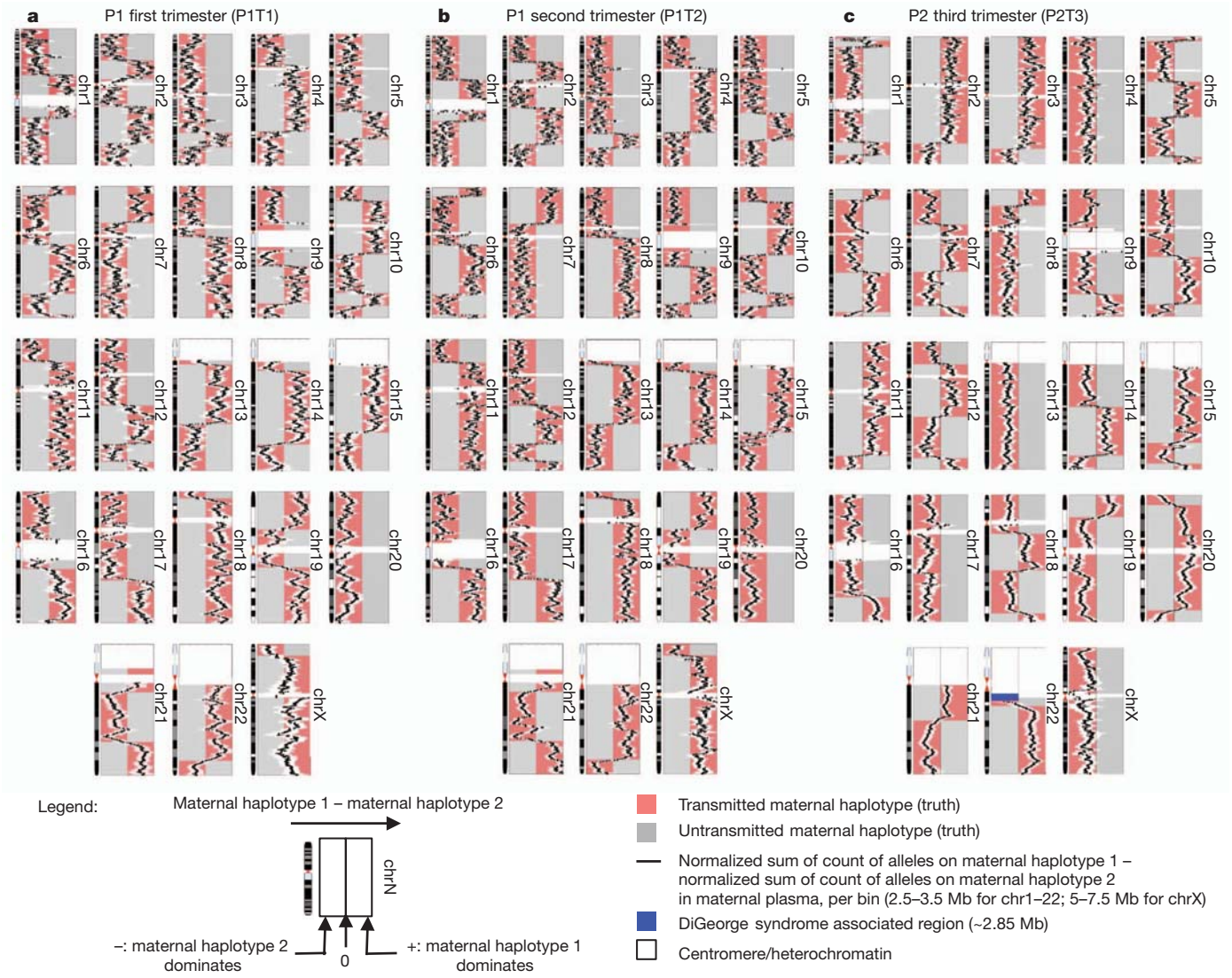


Figure 2 | Non-invasively determining genome-wide fetal inheritance of maternal haplotypes via haplotype counting of maternal plasma DNA with at least 99.8% accuracy over 99.2% of the genome in three maternal plasma samples. **a**–**c**, Each point on a black line represents the relative amount of the two maternal haplotypes evaluated using the markers lying within a bin centred at the point, and is accompanied by a white bar that corresponds to the 95% confidence interval for each measurement in P1 first trimester (**a**), P1 second trimester (**b**) and P2 third trimester (**c**). chr, chromosome. The maternal haplotypes are coloured pink or grey according to the true transmission states, as determined by fetal cord blood genotypes. Over-representation of 'maternal haplotype 2' in P2T3 maternal plasma immediately adjacent to the DiGeorge syndrome associated deletion (blue) indicates fetal inheritance of the deletion, which agrees with fetal cord blood genotype.

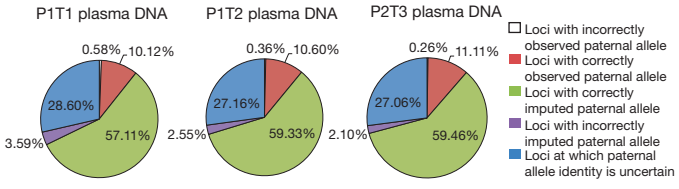


Figure 3 | Reconstruction of paternally inherited chromosomes non-invasively based on imputation using observed non-maternal alleles. The paternally inherited haplotypes were reconstructed by detection of paternal-specific alleles, followed by imputation at linked positions. At the final sequencing depth, ~66–70% of all the paternal-specific alleles were detected at least once. Using those markers, ~70% of the paternally inherited haplotypes were imputed with ~94–97% accuracy. The loci that could not be confidently imputed could in principle be completely determined by deeper sequencing and application of the counting principle directly to the individual alleles at every genomic position.

that lack paternal-specific alleles, or where the paternal alleles are associated with more than one haplotype observed in the population. In principle these regions could be completely determined by deeper sequencing and application of the counting principle directly to the local regions or the individual alleles at every genomic position, as shown below.

Counting alleles at individual loci measures fetal exome

We sought to determine clinically relevant portions of the fetal genome in maternal plasma DNA by applying the counting principle to each allele at all positions in the exome. Because the exome is two orders of magnitude smaller than the genome, less sequencing throughput is required to provide deep sequencing at individual loci and thus allows sensitive and specific detection of clinically relevant and deleterious polymorphisms that were either paternally inherited alleles or *de novo* mutations. We performed exome capture and sequencing on maternal plasma DNA samples of P1 in all three trimesters (Fig. 1 and Supplementary Fig. 9). We obtained a median coverage of 194×, 221× and 631× per position in the exome for the

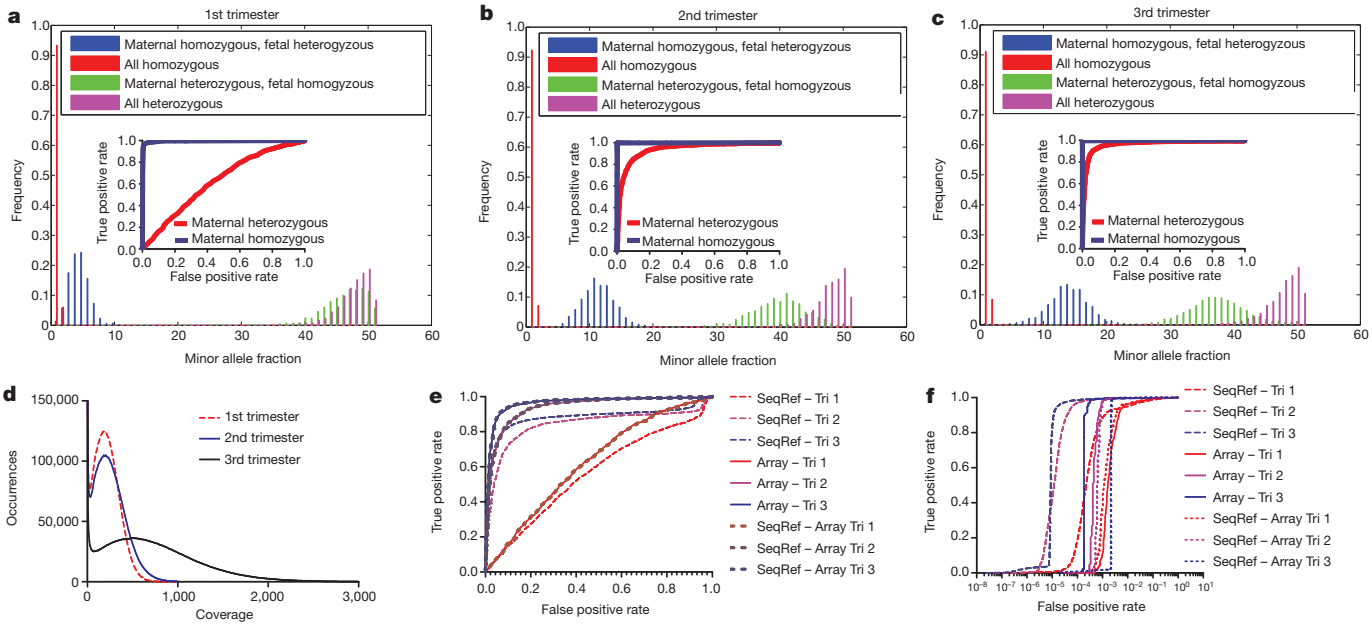


Figure 4 | Exome sequencing of P1 maternal plasma DNA in all three trimesters to determine maternal and fetal genotypes. **a**–**c**, Histograms of minor allele fraction in maternal plasma from first (**a**), second (**b**) and third (**c**) trimesters of P1 at positions that are confidently called in both plasma sequencing data and pure fetal/maternal DNA genotyping data. Insets: Receiver operating characteristic (ROC) curves of positions detecting fetal genotypes differing from maternal genotype when the maternal position is either homozygous or heterozygous. The higher the fetal fraction (~6, 20, 26% for trimester 1, 2, 3, respectively), the more the distributions are separated, and the easier it is to distinguish between the two distributions of fetal genotype. **d**, Histogram of per-position coverage, with bin size of 5. Exome positions >100× are 75%, 78% and 90% respectively for trimester 1, 2, and 3, respectively, and >200× are 48%, 56% and 84%. **e**, **f**, ROCs curves at genomic positions where mother is heterozygous (**e**) or homozygous (**f**), using either sequencing or SNP array of pure DNA as references for maternal and fetal genotypes. 'SeqRef' uses a sequenced reference, 'Array' uses a SNP array, and 'SeqRef-Array' uses a sequenced reference only at positions on a SNP array.

first, second and third trimester, respectively (Fig. 4d). After stringent data filtering to eliminate miscalled paternal-specific alleles due to limited sampling and mis-mapping to the reference genome (Supplementary Fig. 10), 75%, 78% and 90% of all exomic positions in the first, second and third trimester samples, respectively, had >100× coverage and were retained for analysis (Supplementary Table 2).

We calculated minor allele fraction, defined as the second largest nucleotide fraction divided by the sum of the two largest nucleotide fractions, at positions that are confidently called in genotyping data within the exome (Fig. 4a–c) or exome sequencing data (Supplementary Fig. 11–13) of fetal cord blood DNA and pure maternal DNA. In all three trimesters, fetal genotypes could be assigned robustly at loci where the mother is homozygous based on the separation in minor allele fraction at a depth of 200×. Paternal-specific alleles were detected with sensitivity of 96–99.8% at the specificity threshold of 99% (Fig. 4e–f and Table 1). Because the minor allele fraction at loci with paternal-specific alleles is theoretically half of the fetal DNA fraction, we estimated fetal DNA percentage to be 6.6%, 20.1%, 26.3% for the three trimesters, respectively (Supplementary Table 2). For the second and third trimester samples with higher fetal DNA fraction, fetal genotypes could be extracted for most loci at which the mother is heterozygous, as the separation in minor allele fraction for fetal homozygous and fetal heterozygous SNPs was apparent (Fig. 4a–c, e–f). For these loci, the ability to differentiate fetal heterozygosity from homozygosity depended on sequencing depth and fetal DNA fraction (Supplementary Fig. 1).

Table 1 | Exome diagnostic cut-offs and the resulting sensitivity and specificity

			Specificity cut-offs			
			Maternal homozygous		Maternal heterozygous	
	Trimester	Fetal fraction	95%	99%	85%	90%
Sensitivity	1	6%	98%	96%	25%	16%
	2	20%	99.8%	99.8%	89%	85%
	3	26%	99.7%	99.6%	96%	93%

Discussion

The molecular counting methods described here offer a gateway to comprehensive non-invasive prenatal diagnosis of genetic disease. There are substantial ethical issues associated with non-invasive prenatal genome determination, which we have not attempted to address. We will note however that there are numerous clinical scenarios where this approach would be useful. In the first or second trimester, it is possible to test for conditions that are not survivable or lead to medical complications. As technologies for pharmaceutical and surgical intervention improve, it may be possible to develop prenatal treatment or even cures for these congenital conditions.

This is illustrated by our data on P2, who is an individual with DiGeorge syndrome. Haplotyping of the maternal genome identified a ~2.85 Mb deletion on 22q11.1 that is associated with the syndrome on one copy of the maternal chromosome 22 (denoted as 'maternal haplotype 2' in Fig. 2c). Haplotype counting in maternal plasma indicated an over-representation of 'maternal haplotype 2' of the region immediately adjacent to that deletion, indicating fetal inheritance of the DiGeorge syndrome associated deletion (Fig. 2c, deletion indicated in blue). This result was confirmed by quantitative PCR of cord blood DNA (Supplementary Fig. 8). In this clinical scenario, confirmation of the deletion would argue for a fetal echocardiogram and neonatal assessment of calcium levels.

Knowledge of the fetal genotypes obtained in the third trimester enables diagnosis of conditions that would benefit from treatment immediately after delivery; these include metabolic and immunological disorders such as phenylketonuria, galactosaemia, maple syrup urine disease, and severe combined immunodeficiency. Currently, newborns with these conditions suffer as symptoms manifest themselves in the time it takes to determine the proper diagnosis and treatment,

which is often as simple as diet change. In summary, we anticipate that there is no technical barrier and many practical applications to having the entire fetal genome determined non-invasively in clinical settings.

METHODS SUMMARY

Two pregnant subjects (P1 and P2) were recruited with informed consent and approval of the Internal Review Board of Stanford University. Peripheral blood was prospectively obtained at each trimester during the course of pregnancy and post delivery. Direct deterministic phasing (DDP) was performed on three to four single cells obtained from cultures of maternal blood lymphocytes[15]. Cell-free DNA was extracted from maternal plasma during pregnancy and converted into Illumina sequencing libraries using previously established methods[4]. Exome capture was performed on cell-free DNA using SeqCap EZ v2.0 Kit (Roche NimbleGen). Genomic DNA from postpartum maternal blood cells and cord blood cells were assessed by genotyping array (Illumina's HumanOmni1-Quad) and exome sequencing to provide the reference genotypes of the mother and the fetus.

To detect the over-represented parental haplotypes, each chromosome was divided into equally sized bins with sliding window of 100 kb. The bin size was chosen such that the average count was at least that required to overcome counting noise when determining relative representation of the two maternal haplotypes. The relative representation of the maternal haplotypes was calculated using the expression $(N_{p1}/n_{p1} - N_{p2}/n_{p2})$, where N_{pi} is the number of occurrences of markers defining 'maternal haplotype *i*' within the bin counted by sequencing, n_{pi} is the total number of usable markers that define 'maternal haplotype *i*' within the bin. If the expression was positive, maternal haplotype 1 was considered inherited. If the expression was negative, maternal haplotype 2 was considered inherited. Imputation of the allelic identity on unobserved loci was calculated with Impute v1 (ref. 17) using the –haploid option.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 1 March; accepted 23 May 2012.
Published online 4 July 2012.

1. Mandel, P. & Metais, P. Les acides nucléiques du plasma sanguin chez l'homme. *C. R. Acad. Sci. Paris* **142,** 241–243 (1948).
2. Lo, Y. M. *et al.* Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. *Am. J. Hum. Genet.* **62,** 768–775 (1998).
3. Bodurtha, J. & Strauss, J. F. III. Genomics and perinatal care. *N. Engl. J. Med.* **366,** 64–73 (2012).
4. Fan, H. C., Blumenfeld, Y. J., Chitkara, U., Hudgins, L. & Quake, S. R. Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc. Natl Acad. Sci. USA* **105,** 16266–16271 (2008).
5. Sehnert, A. J. *et al.* Optimal detection of fetal chromosomal abnormalities by massively parallel DNA sequencing of cell-free fetal DNA from maternal blood. *Clin. Chem.* **57,** 1042–1049 (2011).
6. Bianchi, D. W. *et al.* Genome-wide fetal aneuploidy detection by maternal plasma DNA sequencing. *Obst. Gynecol.* **119,** 890–901 (2012).
7. Palomaki, G. E. *et al.* DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study. *Genet. Med.* **14,** 296–305 (2012).
8. Palomaki, G. E. *et al.* DNA sequencing of maternal plasma to detect Down syndrome: an international clinical validation study. *Genet. Med.* **13,** 913–920 (2011).
9. Ehrich, M. *et al.* Noninvasive detection of fetal trisomy 21 by sequencing of DNA in maternal blood: a study in a clinical setting. *Am. J. Obstet. Gynecol.* **204,** 205.e1–211.e11 (2011).
10. Chiu, R. W. *et al.* Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study. *Br. Med. J.* **342,** c7401 (2011).
11. Lo, Y. M. *et al.* Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* **2,** 61ra91 (2010).
12. Fan, H. C. & Quake, S. R. In principle method for noninvasive determination of the fetal genome. Preprint at http://precedings.nature.com/documents/5373/version/1 (2010).
13. Macintyre, S. & Sooman, A. Non-paternity and prenatal genetic screening. *Lancet* **338,** 869–871 (1991).
14. Bellis, M. A., Hughes, K., Hughes, S. & Ashton, J. R. Measuring paternal discrepancy and its public health consequences. *J. Epidemiol. Community Health* **59,** 749–754 (2005).
15. Fan, H. C., Wang, J., Potanina, A. & Quake, S. R. Whole-genome molecular haplotyping of single cells. *Nature Biotechnol.* **29,** 51–57 (2011).
16. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
17. Marchini, J. *et al.* A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* **78,** 437–450 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors would like to thank E. Kogut and staff of the Division of Perinatal Genetics and the General Clinical Research Center of Stanford University for coordination of patient recruitment; R. Wong for initial sample processing of clinical samples; N. Neff, G. Mantalas, B. Passarelli and W. Koh for their help in sequencing library preparation and data analysis.

Author Contributions H.C.F., W.G. and S.R.Q. conceived the study. H.C.F., W.G. and J.W. performed experiments. H.C.F., W.G. and J.W. analysed the data. Y.J.B. and Y.Y.E.-S. coordinated patient recruitment. H.C.F., W.G., J.W. and S.R.Q. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.R.Q. (quake@stanford.edu).

METHODS

Prediction of counting depth requirement for determination of over-representation of transmitted maternal haplotypes. Given two distributions of Poisson random variables, one with mean of N, and the other with mean of $N(1-\varepsilon)$, where N is the cumulative sum of the count of all usable markers on the transmitted maternal haplotype, the sampling requirement of N to differentiate the two distributions can be estimated from the following expression, using the normal approximation of the Poisson distribution for large values of N:

$$\frac{N-N(1-\varepsilon)}{\sqrt{N(1-\varepsilon)+N}}=\frac{N\varepsilon}{\sqrt{N(1-\varepsilon)+N}}\geq z_{\alpha}$$

where z_{α} is the z-score associated with the confidence level of α. Thus,

$$N\geq\frac{z_{\alpha}^{2}(2-\varepsilon)}{\varepsilon^{2}}$$

Supplementary Table 1 present the estimated requirement of N for different values of fetal DNA fraction (ε) and level of confidence (α).

Patient samples. Two subjects, referred to as P1 and P2, were recruited to the study under approval of the Internal Review Board of Stanford University. For P1, peripheral blood was obtained during the first, second and third trimesters, and postpartum (Supplementary Table 2). For P2, peripheral blood was obtained during the third trimester and postpartum (Supplementary Table 2). Cord blood was obtained at delivery for both patients.

Whole-genome haplotyping of patient subjects. Postpartum maternal whole blood was collected into sodium heparin coated Vacutainer. Postpartum blood was used in this study because blood samples collected during pregnancy were not cryopreserved on the basis of blood culture requirement. One millilitre of whole blood was cultured with PB Max Karyotyping medium for 4 days. Direct deterministic phasing (DDP)[15] was performed on three to four single cells. Each haplotype was genotyped with Illumina's Omni1-Quad genotyping array. About 92% to 96% of the ~1 million SNPs present on the Omni1Quad BeadChip array (Illumina) (~25% are heterozygous within each individual) were phased (Supplementary Fig. 4), yielding 250–350 heterozygous markers per 3.5-Mb window. In addition to genotyping array analysis, PCR was performed on amplified materials from separated chromosome 22 of P2 to determine which maternal haplotype carried the DiGeorge syndrome associated deletion. Two regions within the deletion were tested, dgs37 and dgs40. Primer sequences are listed in Supplementary Table 3. Other rare SNPs not present on the genotyping array or not linked to loci on the array could also be phased by PCR or sequencing.

Whole-genome genotyping of the study subjects and their infants. Genomic DNA was extracted from 200 µl of postpartum maternal blood and 200 µl cord blood using QIAamp Blood Mini Kit (Qiagen), and subjected to genome-wide genotyping on Illumina's Omni1-Quad genotyping array.

Quantitative PCR confirmation of fetal inheritance of DiGeorge-associated deletion. The inheritance of the maternal haplotype carrying the deletion on chromosome 22q11.1 by the fetus of P2 was independently confirmed by quantitative real-time PCR performed on cord blood genomic DNA. The quantity of an amplicon within the deletion region (Supplementary Table 3) was compared to that of an amplicon on chromosome 1 (EIF2C1). A ratio of ~0.5 indicated that the maternal deletion was inherited.

Extraction of cell-free DNA from maternal plasma. Maternal blood were collected into EDTA coated Vacutainers. Blood was centrifuged at 1,600*g* for 10 min at 4 °C, and the plasma was centrifuged again at 16,000*g* for 10 min at 4 °C to remove residual cells. Cell-free DNA was extracted from plasma using QIAamp Blood Mini Kit (Qiagen) or QIAamp Circulating Nucleic Acid Kit (Qiagen).

Whole-genome shotgun sequencing of cell-free DNA extracted from maternal plasma. DNA was extracted from 1 to 2 ml of plasma, and subsequently converted into Illumina sequencing libraries[4] and quantified by digital PCR[18]. Sequencing was performed on the GAII and the HiSeq instruments (Supplementary Table 2). Sequences were aligned to the human genome (hg19) using CASAVA version 1.7.0. Only alleles called with quality scores >30 were used. In addition, only alleles that match previously reported variants in dbSNP were used for analyses.

Identifying the inherited parental haplotypes. Each chromosome was divided into equally sized bins with sliding window of 100 kb. The bin size was chosen such that the total number of count of markers within the bin was at least that required to overcome counting noise specifically when determining relative representation of the two maternal haplotypes.

The relative representation of the haplotype pairs of each parent was calculated using the expression $(N_{p1}/n_{p1}-N_{p2}/n_{p2})$, where N_{p1} is the number of occurrences of markers defining 'maternal or paternal haplotype 1' within the bin counted by sequencing, n_{p1} is the total number of usable markers that define 'maternal or paternal haplotype 1' within the bin, N_{p2} is the number of occurrences of markers defining 'maternal or paternal haplotype 2' within the bin counted by sequencing, n_{p2} is the total number of usable markers that define 'maternal or paternal haplotype 2' within the bin. If the expression was positive over a continuous region of 5 Mb, parental haplotype 1 was considered as inherited. If the expression was negative over a continuous region of 5 Mb, parental haplotype 2 was considered as inherited. The 95% confidence interval of relative maternal haplotype representation calculated within each bin was estimated by simulating the distribution of reads assuming the count of each maternal haplotype was the mean of a Poisson random variable.

Determining locations of recombination. The true recombination events on the maternally inherited sets of chromosomes were determined by comparing the genotype of the fetus and to the allele on each of the two maternal haplotypes at locations where the fetus is homozygous and the mother is heterozygous. In maternal plasma, a crossover event between the two maternal haplotypes giving rise to the maternally inherited chromosome in the fetus was called if in plasma DNA if two criteria were met: (1) a continuous increase or decrease in the relative representation of haplotype 1 over haplotype 2 (that is, the expression $N_{p1}/n_{p1}-N_{p2}/n_{p2}$), accompanied by a sign change, as one scanned in the direction from the p arm to the q arm of a chromosome; (2) the sign of the expression remained the same for the sliding bins 5 Mb downstream, on the basis of the fact that crossovers are rarely close to each other (positive interference).

Imputation of untyped loci on experimentally measured haplotypes. Imputation was performed using Impute v1 (ref. 17), using the –haploid option. Imputation was performed using August 2010 data from the 1000 Genome Project of the CEU population. For maternal genomes, imputation was based on the ~1 million markers phased by DDP. For paternal haplotypes, imputation was based on non-maternal alleles observed in shotgun sequencing data at locations where mother is observed and predicted based on imputation (>99% confidence) to be homozygous. Only loci with confidence of imputation >99% were considered; the allele identity for the rest were deemed uncertain. The results were compared to the true paternal haplotypes derived on the basis of the comparison of the phased maternal genome and the cord blood genotyping array data. Imputation was performed in 5-Mb segments along each chromosome.

Estimating fetal DNA fraction from maternal plasma sequencing by comparing maternal haplotype representation. Fetal DNA fraction was estimated from the over-representation of one of the maternal haplotypes. Precisely, fetal DNA fraction (ε) was estimated as $2x/(2+x)$, where x is the median absolute value of the expression $(N_{p1}/n_{p1}-N_{p2}/n_{p2})$ for all bins evaluated on either the maternal haplotypes, divided by the average marker density of the two maternal haplotypes.

Exome enrichment from maternal genomic DNA, fetal genomic DNA and cell-free DNA extracted from maternal plasma. Exome capture was performed with the SeqCap EZ v2.0 Kit (Roche NimbleGen) according to manufacturer's protocol with modifications. There are several commercially available exome kits available with varying degrees of coverage for exons, untranslated region, and microRNA regions[19]. We chose the NimbleGen platform due to its ability to capture efficiently on targeted regions and our desire for cost-efficient deep sequencing, but other platforms may perform similarly when sequenced at enough depth.

For exome enriched directly from genomic DNA extracted from maternal blood cells and cord blood, DNA was first sheared using Covaris S220 using the recommended settings for 200-base pair fragments. End repair and dA tailing reactions were cleaned up by QIAquick PCR Purification Kit (Qiagen) whereas ligation and PCR were cleaned by Agencourt Ampure XP beads (Beckman Coulter) at a 1.8× ratio of bead reagent to input volume to discard shorter adaptors, primers and ligation/PCR by-products.

Cell-free DNA extracted from approximately 3, 4, 4 and 2.5 ml of plasma were extracted from P1T1, P1T2, P1T3 and P2, respectively, was used for exome capture. For exome capture from cell-free DNA, sequencing libraries were first prepared following the NEBNext Master Mix 1 Kit (NEB). Extracted DNA was end-repaired and dA-tailed using the NEBNext kit and subsequently cleaned up with QIAquick Nucleotide Removal Kit (Qiagen) in both steps. Ligation to typical Illumina paired-end adaptors was performed at a 1:10 concentration ratio of the initial sample DNA to the adaptors. The first PCR before hybridization was carried for 18 cycles as detailed in the SeqCap protocol. Both ligation and PCR were cleaned up with Agencourt Ampure XP beads as described in the NimbleGen protocol. Prepared non-exome sequencing libraries were incubated with SeqCap kit reagents and the exome-rich sequencing library was amplified for 18 cycles in the second PCR. Libraries were quantified with digital PCR[18].

Analysis of exome sequencing data. Supplementary Fig. 10 outlined the informatics pipeline for analysing exome data. Paired-end sequencing for 100 bases on each end was performed on the HiSeq 2000 (Illumina) using v3

chemistry. Illumina's native software provided image analysis and base calling to provide FASTQ files. Those files were aligned via BWA's 'sampe' function.

Exome sequencing yielded 332, 344 and 930 million aligned reads for first, second and third trimesters, respectively (Supplementary Table 2). Because exome preparation involved more procedural steps and cycles of PCR than whole-genome shotgun sequencing preparation, we imposed a set of filters on the exome data. To remove or at least minimize bias, we opted to remove PCR duplicates on the basis of aligned location with the Picard MarkDuplicates program (the Broad Institute)[20]. In this deduplication procedure, reads with ends aligned to the exact same locations are considered PCR duplicates and amplified from same original single molecule. Deduplication helps substantially reduce bias when using paired-end and sequencing depths exceeding the sample library size. For single end-reads 100 bases long, there is only a maximum unique identification of 200 (for both directions). However, for paired end reads both ends of a DNA fragment are aligned and if fragments lengths are varied equally by 50 bases then the maximum identification library size can be 10,000, which is at least an order of magnitude above the highest coverage seen in this study. In theory it is possible to remove nearly all PCR bias if sequencing is deep enough to discover under-amplified DNA and if the theoretical identification library size is well above the actual molecular library size.

After deduplication, reads were piped through GATK (the Broad Institute) local realigner. Samtools mpileup was used to stack per position counts of different nucleotides within the exome tiles provided by the manufacturer of the SeqCap exome kit. The nucleotide count of each position was analysed against pure fetal and maternal DNA genotyping and sequencing data using custom python and MATLAB code. The minor allele fraction at each position was calculated to be the second largest nucleotide fraction divided by the sum of the two largest nucleotide fractions.

Given that fetal heterozygous genotypes at positions where maternal is homozygous can have a minor allele fraction as low as 1% on the lower end of the distribution, it is important to have more than 100× coverage to avoid classification errors occurring by chance. Beyond 100× coverage, there are also marginal improvements in sensitivity and specificity (Supplementary Fig. 14a). In addition, we filtered out misaligned regions by detecting regions with several excessively high minor allele fractions in close proximity. We filtered out 3–4 positions 40 bases apart with minor allele fractions greater than 1–5% and were able to achieve marked improvement in specificity (Supplementary Fig. 14b). Whereas filtering removes up to 4% of all positions (Supplementary Fig. 14c), it can reduce false positives by an order of magnitude at approximately the same level of sensitivity (Supplementary Fig. 14b).

Fetal DNA fraction was estimated from exome data based on minor allele fraction. The theoretical minor allele fractions are 0 for group 1 SNPs at which both mother and fetus are homozygous, $\varepsilon/2$ for group 2 SNPs of which fetus is heterozygous and mother is homozygous, $1 - \varepsilon/2$ for group 3 SNPs at which fetus is homozygous and mother is heterozygous, and 1/2 for group 4 SNPs at which both mother and fetus are heterozygous, where ε is the fetal DNA fraction. We used the median of the distribution of minor allele fraction for group 2 SNPs to provide an estimate of fetal DNA fraction.

18. White, R. A. III, Blainey, P. C., Fan, H. C. & Quake, S. R. Digital PCR provides sensitive and absolute calibration for high throughput sequencing. *BMC Genomics* **10,** 116 (2009).
19. Clark, M. J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nature Biotechnol.* **29,** 908–914 (2011).
20. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl Acad. Sci. USA* **108,** 9530–9535 (2011).

ARTICLE

doi:10.1038/nature11260

Myocardial infarction accelerates atherosclerosis

Partha Dutta[1]*, Gabriel Courties[1]*, Ying Wei[2], Florian Leuschner[1,3], Rostic Gorbatov[1], Clinton S. Robbins[1], Yoshiko Iwamoto[1], Brian Thompson[1], Alicia L. Carlson[1], Timo Heidt[1], Maulik D. Majmudar[1,4], Felix Lasitschka[5], Martin Etzrodt[1], Peter Waterman[1], Michael T. Waring[6,7], Adam T. Chicoine[6,7], Anja M. van der Laan[8], Hans W. M. Niessen[9], Jan J. Piek[8], Barry B. Rubin[10], Jagdish Butany[11], James R. Stone[1,12], Hugo A. Katus[3], Sabina A. Murphy[13], David A. Morrow[13], Marc S. Sabatine[13], Claudio Vinegoni[1], Michael A. Moskowitz[2], Mikael J. Pittet[1], Peter Libby[4], Charles P. Lin[1], Filip K. Swirski[1], Ralph Weissleder[1,14] & Matthias Nahrendorf[1]

During progression of atherosclerosis, myeloid cells destabilize lipid-rich plaques in the arterial wall and cause their rupture, thus triggering myocardial infarction and stroke. Survivors of acute coronary syndromes have a high risk of recurrent events for unknown reasons. Here we show that the systemic response to ischaemic injury aggravates chronic atherosclerosis. After myocardial infarction or stroke, *Apoe*$^{-/-}$ mice developed larger atherosclerotic lesions with a more advanced morphology. This disease acceleration persisted over many weeks and was associated with markedly increased monocyte recruitment. Seeking the source of surplus monocytes in plaques, we found that myocardial infarction liberated haematopoietic stem and progenitor cells from bone marrow niches via sympathetic nervous system signalling. The progenitors then seeded the spleen, yielding a sustained boost in monocyte production. These observations provide new mechanistic insight into atherogenesis and provide a novel therapeutic opportunity to mitigate disease progression.

Today, survival after a first myocardial infarction (MI) approaches 90%. However, re-infarction occurs commonly and has a high mortality. In a representative trial, new myocardial ischaemia occurred in 54% of patients within the first year after MI[1]. The largest population study so far showed a 17.4% 1-year risk of re-infarction[2]. Conventional wisdom infers that these very high rates of secondary events reflect later stages of linear disease progression. This study tested the alternative hypothesis that a first infarct—triggering a burst of acute systemic inflammation aimed at repair of the injured heart—could accelerate atherosclerosis.

Monocytes infiltrate lesions and, together with their lineage-descendant macrophages, instigate inflammation and deliver proteolytic enzymes that digest extracellular matrix and render atherosclerotic plaques unstable[3–7]. Elevated levels of circulating monocytes provide an expanded pool of inflammatory cells available for recruitment to growing arterial lesions, potentially promoting plaque rupture. Leukocytosis after MI predicts an increased risk of re-infarction and death[8,9]. During acute MI, blood monocyte levels spike, and these cells accumulate in the evolving myocardial wound[10,11]. Thus, the organism experiences an acute inflammatory event (for example, MI) superimposed on a pre-existing chronic inflammatory disease (atherosclerosis), both of which involve the same myeloid cell type. Given the frequency of re-infarction, we investigated whether acute myocardial injury accelerates pre-existing chronic atherosclerosis.

We found that in *Apoe*$^{-/-}$ mice with atherosclerosis, MI increased plaque size and induced a 'vulnerable' lesion morphology with higher inflammatory cell content and protease activity, fuelled by persistently increased myeloid cell flux to atherosclerotic sites. Earlier clinical studies described an increase of haematopoietic stem and progenitor cells (HSPCs) in the circulation of patients shortly after MI[12]. We thus proposed that release of these progenitors may increase the availability of monocytes. We found that in response to heightened sympathetic nervous system (SNS) activity—provoked by pain, anxiety and heart failure in patients with MI—HSPCs departed bone marrow niches and produced prolonged amplified extramedullary monocytopoiesis in mice after coronary ligation.

MI accelerates atherosclerosis

Proteases, including metalloproteinases and cysteinyl cathepsins, can catabolize the extracellular matrix of the plaque's fibrous cap and render it prone to rupture[13,14]. Therefore, protease activity may serve as a marker in mice of processes associated with lesion vulnerability in humans[15]. To test the hypothesis that MI changes the course of atherosclerotic disease, we serially imaged protease activity in aortic plaques of *Apoe*$^{-/-}$ mice, before and 3 weeks after coronary ligation, using hybrid fluorescence molecular tomography–X-ray computed tomography (FMT–CT)[16]. Imaging showed a sharp increase of plaque protease activity within 3 weeks after MI (Fig. 1a, b). In parallel,

[1]Center for Systems Biology, Massachusetts General Hospital and Harvard Medical School, Simches Research Building, 185 Cambridge Street, Boston, Massachusetts 02114, USA. [2]Stroke and Neurovascular Regulation Laboratory, Departments of Radiology and Neurology, Massachusetts General Hospital/Harvard Medical School, 149 13th Street, Charlestown, Massachusetts 02129, USA. [3]Department of Cardiology, Medical University Hospital Heidelberg, Im Neuenheimer Feld 410, D-69120 Heidelberg, Germany. [4]Cardiovascular Division, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [5]Institute of Pathology, University Hospital Heidelberg, Im Neuenheimer Feld 220/221, 69120 Heidelberg, Germany. [6]The Ragon Institute of MGH, MIT and Harvard at Massachusetts General Hospital, Charlestown, Massachusetts 02129, USA. [7]Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA. [8]Department of Cardiology, Academic Medical Center, University of Amsterdam, Meibergdreef 9, 1105 AZ, Amsterdam, the Netherlands. [9]Department of Pathology and Cardiac Surgery, ICaR-VU, VU University Medical Center, De Boelelaan 1117, 1081 HV Amsterdam, the Netherlands. [10]Division of Vascular Surgery, Peter Munk Cardiac Centre, Toronto General Hospital, University of Toronto, Toronto, Ontario M5G-2C4, Canada. [11]Department of Pathology, Peter Munk Cardiac Centre, University of Toronto, Toronto, Ontario M5G-2C4, Canada. [12]Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA. [13]TIMI Study Group, Department of Medicine, Cardiovascular Division, Brigham and Women's Hospital, Boston, Massachusetts 02145, USA. [14]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA.
*These authors contributed equally to this work.

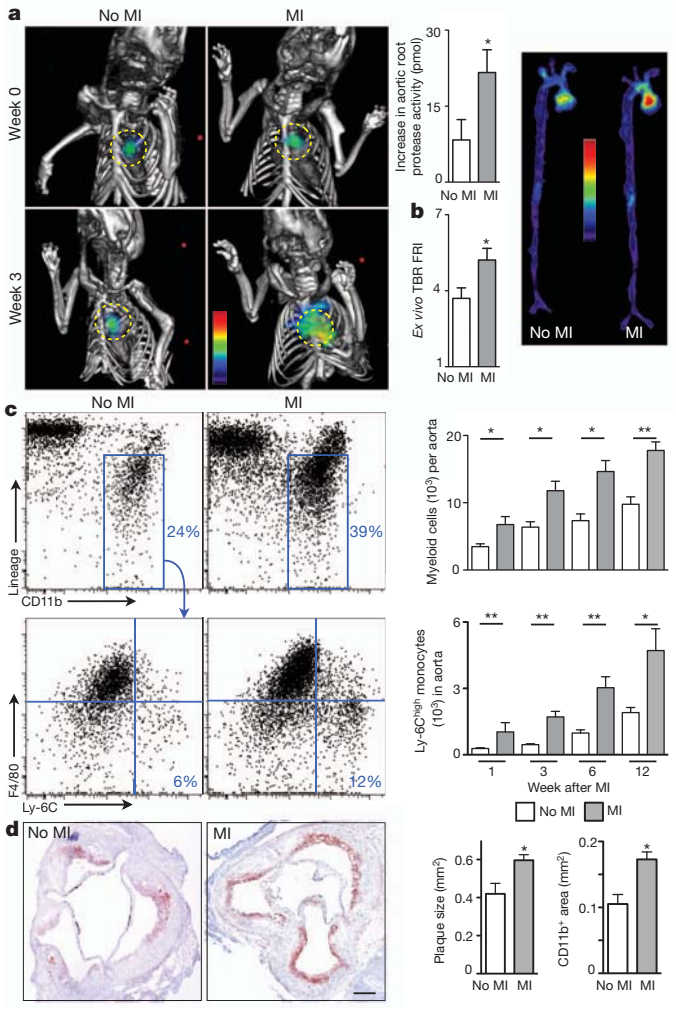


Figure 1 | Increased inflammation in atherosclerotic plaques after MI. **a**, Protease activity was determined by FMT–CT before and 3 weeks after MI. Circles indicate aortic root ($n = 10$ per group). **b**, Protease activity in excised aortae determined by fluorescence reflectance imaging (FRI), expressed as target to background ratio (TBR; $n = 10$ per group). **c**, Flow cytometric quantification of myeloid cells and Ly-6C^{high} monocytes in aorta ($n = 5$–9 per group). Dot plots 3 weeks after MI are shown. **d**, CD11b staining and lesion size ($n = 9$–10 per group). Scale bar represents 150 μm. Data are shown as mean ± s.e.m. $^*P < 0.05$, $^{**}P < 0.01$.

expression of the inflammatory cytokine interleukin-6 (*Il6*), *Mmp9*, myeloperoxidase and *Ly-6C* (also known as *Ly6c1*) increased in atherosclerotic plaques (Supplementary Fig. 1). The number of monocytes and macrophages per aorta increased, particularly the inflammatory Ly-6C^{high} monocyte subset (Fig. 1c). Plaque monocyte content also increased in *Apoe*$^{-/-}$ mice without MI, reflecting the natural course of disease in these animals[17,18]. Yet innate immune cell accumulation accelerated distinctively after MI, as indicated by the significantly greater slope obtained when fitting the number of Ly-6C^{high} monocytes in the aorta over time (Supplementary Fig. 2). Neutrophil presence in atheromata also increased (Supplementary Fig. 3) whereas mast cells did not (Supplementary Fig. 4). Histological analysis affirmed increased accumulation of CD11b$^+$ myeloid cells and larger lesion size after MI (Fig. 1d). The thickness of the fibrous cap decreased, covering larger necrotic cores (Supplementary Fig. 5). Ly-6C^{high} monocytes isolated from atherosclerotic lesions exhibited higher levels of messenger RNAs encoding inflammatory genes. *Il1b* and cathepsin B were expressed at higher levels 3 weeks after MI, whereas arginase (Arg1) and TGF-β, markers associated with alternatively activated macrophages, were expressed at lower levels (Supplementary Fig. 6). Monocyte numbers in the blood and spleen increased consistently for up to 3 months after coronary ligation (Supplementary Fig. 7) but were unaltered in the bone marrow (Supplementary Fig. 8).

Extramedullary monocytopoiesis after MI

Because the spleen has the ability to host extramedullary haematopoiesis[19–21], we measured splenic monocyte progenitor content in mice after MI. Haematopoietic progenitor cell numbers in the spleen increased after MI (Fig. 2 and Supplementary Fig. 9) but not in the bone marrow (Supplementary Fig. 10). Proliferation of progenitors doubled in the spleen (Supplementary Fig. 11). In patients who died after an acute MI, we found increased numbers of c-kit$^+$ cells in the spleen, some of which co-localized with the proliferation marker Ki-67 (Supplementary Fig. 12).

When we splenectomized mice at the time of MI, atherosclerosis did not accelerate (Supplementary Fig. 13). The number of progenitor cells in liver tissue after MI was much lower than in the spleen; however, splenectomy increased progenitor cell presence in the liver 4 days after MI (Supplementary Fig. 14). We concluded that the infarct-induced monocytosis resulted primarily from augmented production in the spleen, but that other extramedullary sites may contribute[22]. This observation raised the question whether monocytes of splenic and bone marrow origin differ qualitatively. Surprisingly, Ly-6C^{high} monocytes isolated from the spleen or bone marrow on day 4 after MI had significantly different mRNA levels in 11 of the 32 genes assessed (Supplementary Fig. 15). For instance, *Il1b* and cathepsin B mRNA levels were 60- and 6-fold higher in inflammatory monocytes isolated from the spleen, matching the increased expression of these genes in Ly-6C^{high} monocytes isolated from atherosclerotic plaques after MI (Supplementary Fig. 6). Therefore, post-MI extramedullary myelopoiesis may not only increase the availability of inflammatory cells but also change their functional program. To test whether another form of acute tissue injury prevalent in atherosclerotic patients would accelerate splenic myelopoiesis, we analysed *Apoe*$^{-/-}$ mice 6 weeks after ischaemic stroke. The number of myeloid cells and Ly-6C^{high} monocytes in atherosclerotic plaques increased after stroke, in parallel with expanded splenic monocytopoiesis (Supplementary Fig. 16).

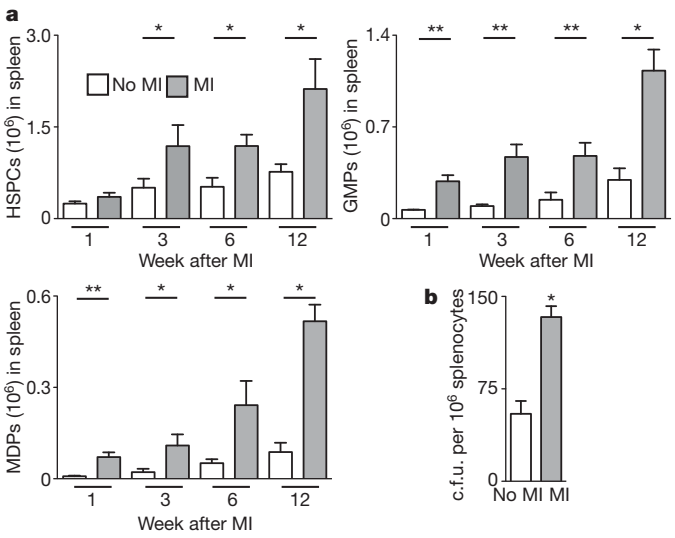


Figure 2 | Elevated levels of progenitor cells in the spleen of *Apoe*$^{-/-}$ mice after MI. **a**, Quantification of HSPCs, MDPs and GMPs at different time points after MI ($n = 3$–15 per group). The gating strategy is shown in Supplementary Fig. 10. **b**, Number of colony-forming units (c.f.u.). Data are shown as mean ± s.e.m. $^*P < 0.05$, $^{**}P < 0.01$.

Bone marrow HSPC release after MI

As granulocyte macrophage progenitors (GMPs) and macrophage dendritic cell progenitors (MDPs) have a limited self-renewal capacity[23,24], we tested whether upstream progenitors released from their bone marrow niches sustain the splenic proliferative activity after MI. Indeed, blood levels of HSPCs increased 2-, 7- and 24-fold at 6, 48 and 96 h after MI, respectively (Fig. 3a). The number of splenic Flk2$^-$ HSPCs increased markedly after MI (Supplementary Fig. 17). This mobilization of upstream HSPCs with high capacity for self-renewal probably explains the long-term boost in splenic monocyte production in *Apoe*$^{-/-}$ mice after MI.

Anxiety, pain and impaired left ventricular function during MI can all activate the SNS. Accordingly, levels of tyrosine hydroxylase, the rate-limiting enzyme for production of noradrenaline in sympathetic fibres[25], increased in the bone marrow of mice after MI and hence indicated a higher sympathetic tone (Fig. 3b). SNS activity may liberate haematopoietic stem cells from their niches by signalling through the β_3-adrenoceptor[26]. Nestin$^+$ mesenchymal stem cells express this receptor, which regulates the production of stem cell retention factors[27]. Because acute MI raises blood progenitor levels in patients[12], we investigated whether SNS activity causes the release of HSPCs from the bone marrow after MI. Blood HSPCs decreased by 100, 75 and 50% at 6, 48 and 96 h after MI in mice treated with a β_3-adrenoceptor antagonist (Fig. 3a). The stem cell retention factor *Cxcl12*, angiopoietin and stem cell factor (*Scf*; also known as *Kitl*)[28] underwent similar regulation (Fig. 3c). Levels of the adhesion molecule *Vcam1*, which also retains HSPCs in the bone marrow, decreased after MI but did not change after β_3-adrenoceptor blocker administration (Fig. 3c). These data indicate that increased sympathetic tone after MI causes withdrawal of stem cell retention factors by β_3-adrenoceptor-expressing niche cells.

Treatment with a β_3-adrenoceptor blocker reduced splenic accumulation of progenitors in wild-type mice shortly after MI (Supplementary Fig. 18) and consequently diminished their output of myeloid cells (Supplementary Fig. 19). In *Apoe*$^{-/-}$ mice 3 weeks after MI, β_3-blocker treatment reduced the number of GMPs and their progeny in the spleen and blood (Supplementary Fig. 20). Retrospective analysis of a clinical trial[29] revealed that prior β-blocker therapy was associated with a reduction in monocytes after an acute coronary syndrome (Supplementary Table 1). The mechanism that led to this decrease is unclear, also because some clinically used β-blockers have a lower affinity for the β_3-adrenoceptor subtype[30]; however, these associative data show an interesting parallel to our findings in mice.

In *Apoe*$^{-/-}$ mice after MI, β_3-blocker treatment lowered protease activity, myeloid cell content, and mRNA levels of inflammatory cytokines in the plaque (Supplementary Fig. 21). When we adoptively transferred GFP$^+$ GMPs to wild-type mice with MI, β_3-blocker treatment did not alter their splenic differentiation (Supplementary Fig. 22). Sympathetic denervation with 6-hydroxydopamine (6-OHDA)[26,31] increased bone marrow mRNA levels of the stem cell retention factor *Cxcl12*, reduced levels of HSPCs in blood, decreased circulating monocyte levels, and attenuated the accumulation of myeloid cells in atherosclerotic lesions (Supplementary Fig. 23). Combination of β_3 blockade and splenectomy showed no additive effects (Supplementary Fig. 24). Neither MI nor β_3 blockade changed blood cholesterol and high-density lipoprotein levels (Supplementary Fig. 25).

Intravital microscopy of HSPC departure

We adoptively transferred lineage$^-$ c-kit$^+$ Sca-1$^+$ Flk2$^-$ (Sca1 also known as Ly6a) HSPCs labelled with a fluorescent membrane dye (DiD) to examine their release with serial intravital microscopy

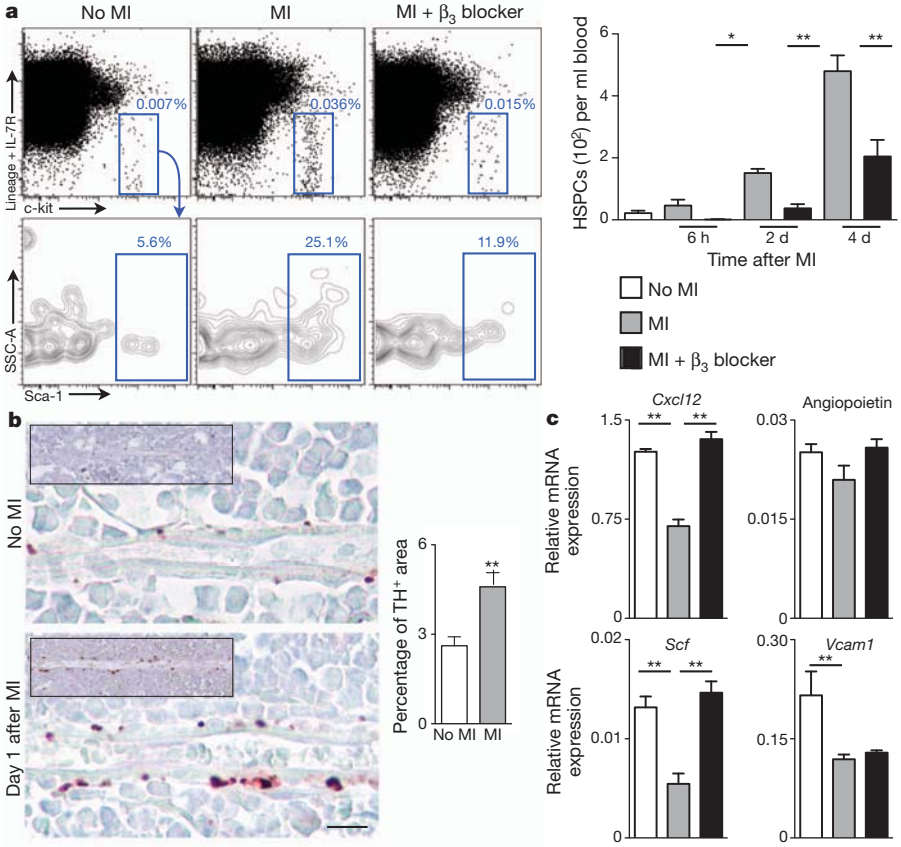


Figure 3 | β_3-Adrenoceptor-mediated progenitor release after MI. **a**, Flow cytometric analyses of HSPCs in blood of C57BL/6 mice ($n = 6$–11 per group). **b**, Immunostaining for tyrosine hydroxylase (TH). Scale bar represents 10 μm. Insets depict low-magnification overview. Bar graph shows quantification of TH$^+$ area ($n = 5$ per group). **c**, Expression of HSPC retention factors (relative to GAPDH) in the bone marrow of C57BL/6 mice on day 4 after MI ($n = 8$ per group). Data are shown as mean ± s.e.m. $^*P < 0.05$, $^{**}P < 0.01$.

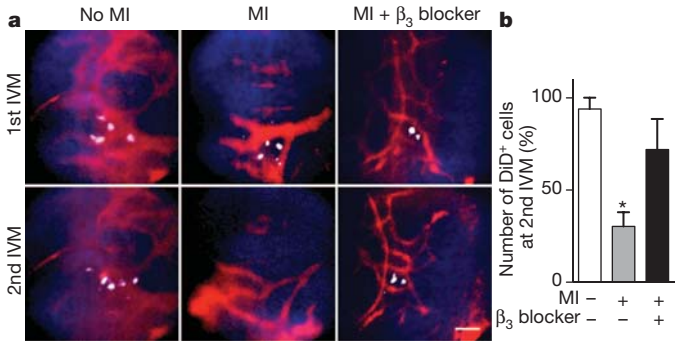


Figure 4 | Serial intravital imaging of progenitor release from the bone marrow. **a**, DiD-labelled-HSPC Flk2$^-$ cells were imaged in the skull bone marrow by IVM before and then again 4 days after MI. DiD–labelled HSPCs are white, blood pool is red, and bone is blue. SSC-A, side scatter. Scale bar represents 50 μm. **b**, Change of HSPC presence between first and second IVM session ($n = 3$ per group). Data are shown as mean ± s.e.m. *$P < 0.05$.

(IVM)[32]. DiD$^+$ cells were quantified after they had settled into the bone marrow, and then again 4 days after MI. Concomitant with the post-MI increase of progenitors in circulation, 52% of cells that were present during the first imaging session departed from the bone marrow, which was inhibited by the β_3-adrenoceptor antagonist (Fig. 4). Post-imaging flow cytometry corroborated the trafficking of DiD$^+$ cells (Supplementary Fig. 26). We next investigated the relocation of bone marrow cells to the spleen directly. Lineage$^-$ c-kit$^+$ Sca-1$^+$ Flk2$^-$ HSPCs were harvested from CD45.2$^+$ donors and labelled with a photoconvertible dye before transfer into CD45.1$^+$ recipients. These cells engrafted into the skull bone marrow, where we photoconverted them with laser illumination. Only if mice underwent coronary ligation, photoconverted CD45.2$^+$ 4′,6-diamidino-2-phenylindole (DAPI)$^+$ cells were detected in splenic cell suspensions 4 days later (Supplementary Fig. 27).

Splenic HSPC engraftment after MI

Finally, we investigated the mechanisms of splenic progenitor seeding. The mRNA levels of *Scf* increased in splenic tissue after MI in parallel with the number of SCF$^+$ cells in splenic sections (Fig. 5a, b). Antibody neutralization of SCF decreased retention of adoptively transferred HSPCs in the spleen and proliferation of host HSPCs (Fig. 5c, d). Co-localization studies identified CD31$^+$ and occasionally nestin$^+$ cells (Supplementary Fig. 28a, b) as a source of SCF, in agreement with a recent report on the role of SCF in the splenic niche during the steady state[33]. We found adoptively transferred DiD$^+$ HSPCs cells in close vicinity to CD31$^+$ cells (Supplementary Fig. 28c). Neutralization of VLA-4 (also known as Itga4), an integrin involved in stem cell retention[34,35], reduced the number of adoptively transferred HSPCs in the spleen after MI, but not in the steady state (Supplementary Fig. 29).

Discussion

We have shown that acute MI or stroke increases inflammation in atherosclerotic plaques at a distance. After an ischaemic event, atherosclerotic plaques grew faster and displayed higher protease activity. We identified an increased supply of innate immune cells as a driving force for this phenomenon. On a systems level, pre-existing chronic inflammation flared when mice experienced an additional acute inflammatory stimulus. Increased SNS activity after MI released upstream progenitors from bone marrow niches. On the receiving end, the spleen hosted these cells by increasing SCF production, leading to amplified extramedullary myelopoiesis (Fig. 5e). The pro-inflammatory changes in atherosclerotic plaques persisted for several months.

The evolutionary benefit of outsourcing myelopoiesis from the bone marrow may involve the protection of steady state 'housekeeping' in this confined compartment. Unlike the bone marrow, the spleen is an organ that can rapidly expand in size. In the event of increased leukocyte need after acute injury, the myelopoietic system may proliferate in extramedullary sites to protect quiescent stem cells and to ensure unimpeded production of red cells, platelets and lymphocytes in the bone marrow.

Despite growing understanding of the chronic inflammatory nature of atherosclerosis[3,6,7], specific anti-inflammatory therapy has yet to materialize. Given the central role of myeloid cells in disease promotion and their rapid turnover in inflamed tissue, interrupting the monocyte supply chain may attenuate atherosclerosis. In this case, SNS inhibition abrogated stem cell release from the bone marrow. Because the regulation of progenitor cell migration is multifactorial[35], there are other targets along this pathway that await exploration, including chemokine receptors and cytokines involved in stem cell activation. In addition, the innate immune response unleashed by acute ischaemic injury may also change the 'fluid phase' of blood by augmenting circulating acute phase reactants such as fibrinogen and

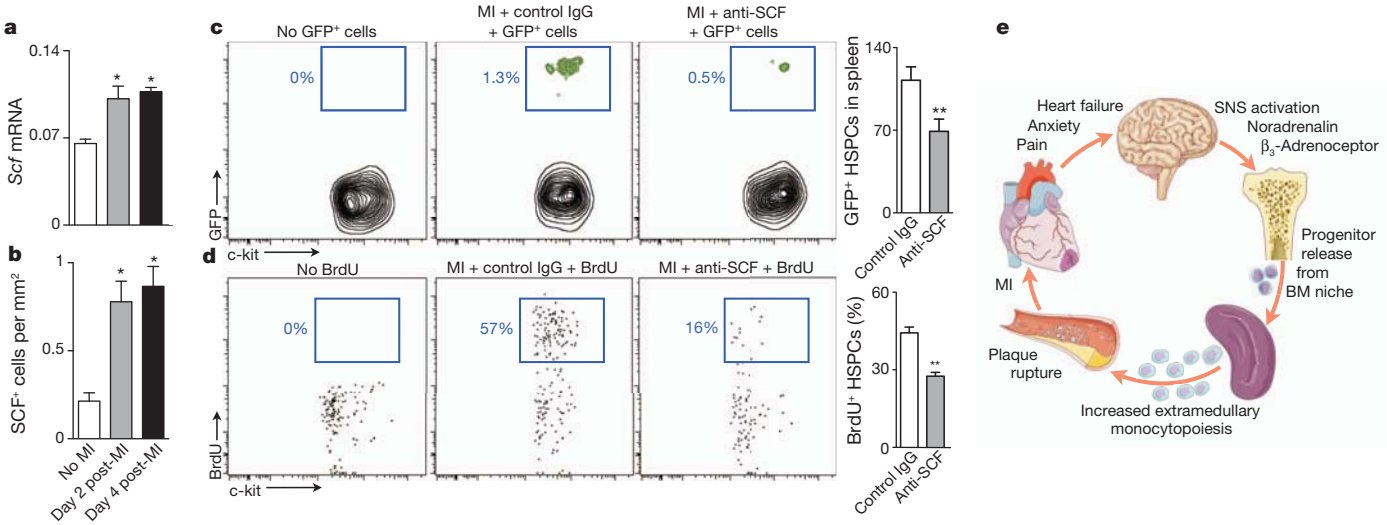


Figure 5 | Splenic progenitor engraftment after MI. **a**, Quantitative polymerase chain reaction of SCF in spleen ($n = 5$–6 per group). **b**, Number of SCF$^+$ cells in spleen of C57BL/6 mice 4 days after MI as determined by immunofluorescence. **c**, Enumeration of adoptively transferred GFP$^+$ HSPCs on day 4 after MI ($n = 8$ per group). **d**, Proliferation of endogenous HSPCs determined by BrdU incorporation ($n = 8$ per group). **e**, Paradigm. BM, bone marrow. Data are shown as mean ± s.e.m. *$P < 0.05$, **$P < 0.01$.

plasminogen activator inhibitor 1, factors that promote thrombosis and counter endogenous fibrinolysis[36]. Our study suggests that patients with an ischaemic complication of atherosclerosis experience a particularly vulnerable disease phase, and that interventions aimed at progenitors of innate immune cells could affect long-term outcomes.

METHODS SUMMARY

Wild-type C57BL/6J, C57BL/6.SJL, C57BL/6-Tg(UBC-GFP)30Scha/J and B6.129P2-Apoe^{tm1Unc}/J mice were used in these studies, which were approved by the Subcommittee on Animal Research Care at Massachusetts General Hospital. The patient studies were conducted in accordance with the Declaration of Helsinki. The studies were approved by the Research Committee of the Department of Pathology of the VUmc and by the Ethikkommission Heidelberg University. Detailed procedures are available in Supplementary Information.

Received 12 December 2011; accepted 25 May 2012.
Published online 27 June 2012.

1. Goldstein, J. A. *et al.* Multiple complex coronary plaques in patients with acute myocardial infarction. *N. Engl. J. Med.* **343,** 915–922 (2000).
2. Milonas, C. *et al.* Effect of angiotensin-converting enzyme inhibition on one-year mortality and frequency of repeat acute myocardial infarction in patients with acute myocardial infarction. *Am. J. Cardiol.* **105,** 1229–1234 (2010).
3. Libby, P., Ridker, P. M. & Hansson, G. K. Progress and challenges in translating the biology of atherosclerosis. *Nature* **473,** 317–325 (2011).
4. Weber, C. & Noels, H. Atherosclerosis: current pathogenesis and therapeutic options. *Nature Med.* **17,** 1410–1422 (2011).
5. Randolph, G. J. The fate of monocytes in atherosclerosis. *J. Thromb. Haemost.* **7** (suppl. 1), 28–30 (2009).
6. Charo, I. F. & Ransohoff, R. M. The many roles of chemokines and chemokine receptors in inflammation. *N. Engl. J. Med.* **354,** 610–621 (2006).
7. Galkina, E. & Ley, K. Immune and inflammatory mechanisms of atherosclerosis. *Annu. Rev. Immunol.* **27,** 165–197 (2009).
8. Ernst, E., Hammerschmidt, D. E., Bagge, U., Matrai, A. & Dormandy, J. A. Leukocytes and the risk of ischemic diseases. *J. Am. Med. Assoc.* **257,** 2318–2324 (1987).
9. Sabatine, M. S. *et al.* Relationship between baseline white blood cell count and degree of coronary artery disease and mortality in patients with acute coronary syndromes: a TACTICS-TIMI 18 substudy. *J. Am. Coll. Cardiol.* **40,** 1761–1768 (2002).
10. Nahrendorf, M. *et al.* The healing myocardium sequentially mobilizes two monocyte subsets with divergent and complementary functions. *J. Exp. Med.* **204,** 3037–3047 (2007).
11. Nahrendorf, M., Pittet, M. J. & Swirski, F. K. Monocytes: protagonists of infarct inflammation and repair after myocardial infarction. *Circulation* **121,** 2437–2445 (2010).
12. Massa, M. *et al.* Increased circulating hematopoietic and endothelial progenitor cells in the early phase of acute myocardial infarction. *Blood* **105,** 199–206 (2005).
13. Galis, Z. S., Sukhova, G. K., Lark, M. W. & Libby, P. Increased expression of matrix metalloproteinases and matrix degrading activity in vulnerable regions of human atherosclerotic plaques. *J. Clin. Invest.* **94,** 2493–2503 (1994).
14. Libby, P. Inflammation in atherosclerosis. *Nature* **420,** 868–874 (2002).
15. Chen, J. *et al. In vivo* imaging of proteolytic activity in atherosclerosis. *Circulation* **105,** 2766–2771 (2002).
16. Nahrendorf, M. *et al.* Hybrid *in vivo* FMT-CT imaging of protease activity in atherosclerosis with customized nanosensors. *Arterioscler. Thromb. Vasc. Biol.* **29,** 1444–1451 (2009).
17. Tacke, F. *et al.* Monocyte subsets differentially employ CCR2, CCR5, and CX3CR1 to accumulate within atherosclerotic plaques. *J. Clin. Invest.* **117,** 185–194 (2007).
18. Swirski, F. K. *et al.* Ly-6C^{hi} monocytes dominate hypercholesterolemia-associated monocytosis and give rise to macrophages in atheromata. *J. Clin. Invest.* **117,** 195–205 (2007).
19. Robbins, C. S. *et al.* Extramedullary hematopoiesis generates Ly-6C^{high} monocytes that infiltrate atherosclerotic lesions. *Circulation* **125,** 364–374 (2012).
20. Leuschner, F. *et al.* Rapid monocyte kinetics in acute myocardial infarction are sustained by extramedullary monocytopoiesis. *J. Exp. Med.* **209,** 123–137 (2012).
21. Swirski, F. K. *et al.* Identification of splenic reservoir monocytes and their deployment to inflammatory sites. *Science* **325,** 612–616 (2009).
22. Psaltis, P. J. *et al.* Identification of a monocyte-predisposed hierarchy of hematopoietic progenitor cells in the adventitia of postnatal murine aorta. *Circulation* **125,** 592–603 (2012).
23. Kondo, M. *et al.* Biology of hematopoietic stem cells and progenitors: implications for clinical application. *Annu. Rev. Immunol.* **21,** 759–806 (2003).
24. Geissmann, F. *et al.* Development of monocytes, macrophages, and dendritic cells. *Science* **327,** 656–661 (2010).
25. Zigmond, R. E. & Ben-Ari, Y. Electrical stimulation of preganglionic nerve increases tyrosine hydroxylase activity in sympathetic ganglia. *Proc. Natl Acad. Sci. USA* **74,** 3078–3080 (1977).
26. Katayama, Y. *et al.* Signals from the sympathetic nervous system regulate hematopoietic stem cell egress from bone marrow. *Cell* **124,** 407–421 (2006).
27. Méndez-Ferrer, S. *et al.* Mesenchymal and haematopoietic stem cells form a unique bone marrow niche. *Nature* **466,** 829–834 (2010).
28. Méndez-Ferrer, S., Lucas, D., Battista, M. & Frenette, P. S. Haematopoietic stem cell release is regulated by circadian oscillations. *Nature* **452,** 442–447 (2008).
29. Cannon, C. P. *et al.* Intensive versus moderate lipid lowering with statins after acute coronary syndromes. *N. Engl. J. Med.* **350,** 1495–1504 (2004).
30. Hoffmann, C., Leitz, M. R., Oberdorf-Maass, S., Lohse, M. J. & Klotz, K. N. Comparative pharmacology of human β-adrenergic receptor subtypes—characterization of stably transfected receptors in CHO cells. *Naunyn Schmiedebergs Arch. Pharmacol.* **369,** 151–159 (2004).
31. Kruszewska, B., Felten, S. Y. & Moynihan, J. A. B. Alterations in cytokine and antibody production following chemical sympathectomy in two strains of mice. *J. Immunol.* **155,** 4613–4620 (1995).
32. Lo Celso, C. *et al.* Live-animal tracking of individual haematopoietic stem/progenitor cells in their niche. *Nature* **457,** 92–96 (2009).
33. Ding, L., Saunders, T. L., Enikolopov, G. & Morrison, S. J. Endothelial and perivascular cells maintain haematopoietic stem cells. *Nature* **481,** 457–462 (2012).
34. Williams, D. A., Rios, M., Stephens, C. & Patel, V. P. Fibronectin and VLA-4 in haematopoietic stem cell–microenvironment interactions. *Nature* **352,** 438–441 (1991).
35. Lo Celso, C. & Scadden, D. T. The haematopoietic stem cell niche at a glance. *J. Cell Sci.* **124,** 3529–3535 (2011).
36. Libby, P. & Theroux, P. Pathophysiology of coronary artery disease. *Circulation* **111,** 3481–3488 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the CSB Mouse Imaging Program (J. Truelove, D. Jeon, J. Donahoe, B. Marinelli) and K. Naxerova for helpful discussions. This work was funded by grants from the National Institute of Health R01-HL096576, R01-HL095629 (M.N.); R01-EB006432, T32-CA79443, P50-CA086355 (R.W.). F.L. was funded in part by Deutsche Forschungsgemeinschaft SFB 938/Z2. Fig. 5e was produced using Servier Medical Art (http://www.servier.com).

Author Contributions P.D. and G.C. performed experiments, collected and analysed the data, and contributed to writing the manuscript, R.G. did surgeries and performed experiments, Y.W., F.Le., R.G., C.S.R., Y.I., B.T., A.L.C., T.H., M.D.M., F.La., M.E., P.W., M.T.W., A.T.C., A.M.v.d.L., H.W.M.N., J.J.P., B.B.R., J.B., J.R.S., H.A.K., C.V., S.A.M., D.A.M. and M.S.S. performed experiments, collected, analysed and discussed data, M.A.M., M.J.P., P.L., C.P.L., F.K.S. and R.W. conceived experiments and discussed strategy and results; M.N. designed and managed the study and wrote the manuscript, which was edited and approved by all co-authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.N. (mnahrendorf@mgh.harvard.edu) or R.W. (rweissleder@mgh.harvard.edu).

ARTICLE

doi:10.1038/nature11252

Comprehensive molecular characterization of human colon and rectal cancer

The Cancer Genome Atlas Network*

To characterize somatic alterations in colorectal carcinoma, we conducted a genome-scale analysis of 276 samples, analysing exome sequence, DNA copy number, promoter methylation and messenger RNA and microRNA expression. A subset of these samples (97) underwent low-depth-of-coverage whole-genome sequencing. In total, 16% of colorectal carcinomas were found to be hypermutated: three-quarters of these had the expected high microsatellite instability, usually with hypermethylation and *MLH1* silencing, and one-quarter had somatic mismatch-repair gene and polymerase ε (*POLE*) mutations. Excluding the hypermutated cancers, colon and rectum cancers were found to have considerably similar patterns of genomic alteration. Twenty-four genes were significantly mutated, and in addition to the expected *APC*, *TP53*, *SMAD4*, *PIK3CA* and *KRAS* mutations, we found frequent mutations in *ARID1A*, *SOX9* and *FAM123B*. Recurrent copy-number alterations include potentially drug-targetable amplifications of *ERBB2* and newly discovered amplification of *IGF2*. Recurrent chromosomal translocations include the fusion of *NAV2* and WNT pathway member *TCF7L1*. Integrative analyses suggest new markers for aggressive colorectal carcinoma and an important role for *MYC*-directed transcriptional activation and repression.

The Cancer Genome Atlas project plans to profile genomic changes in 20 different cancer types and has so far published results on two cancer types[1,2]. We now present results from multidimensional analyses of human colorectal carcinoma (CRC).

CRC is an important contributor to cancer mortality and morbidity. The distinction between the colon and the rectum is largely anatomical, but it has both surgical and radiotherapeutic management implications and it may have an impact on prognosis. Most investigators divide CRC biologically into those with microsatellite instability (MSI; located primarily in the right colon and frequently associated with the CpG island methylator phenotype (CIMP) and hyper-mutation) and those that are microsatellite stable but chromosomally unstable.

A rich history of investigations (for a review see ref. 3) has uncovered several critical genes and pathways important in the initiation and progression of CRC (ref. 3). These include the WNT, RAS−MAPK, PI3K, TGF-β, P53 and DNA mismatch-repair pathways. Large-scale sequencing analyses[4–6] have identified numerous recurrently mutated genes and a recurrent chromosomal translocation. Despite this background, we have not had a fully integrated view of the genetic and genomic changes and their significance for colorectal tumorigenesis. Further insight into these changes may enable deeper understanding of the pathophysiology of CRC and may identify potential therapeutic targets.

Results

Tumour and normal pairs were analysed by different platforms. The specific numbers of samples analysed by each platform are shown in Supplementary Table 1.

Exome-sequence analysis

To define the mutational spectrum, we performed exome capture DNA sequencing on 224 tumour and normal pairs (all mutations are listed in Supplementary Table 2). Sequencing achieved >20-fold coverage of at least 80% of targeted exons. The somatic mutation rates varied considerably among the samples. Some had mutation rates of <1 per 10^6 bases, whereas a few had mutations rates of >100 per 10^6. We separated cases (84%) with a mutation rate of <8.24 per 10^6 (median number of non-silent mutations, 58) and those with mutation rates of >12 per 10^6 (median number of total mutations, 728), which we designated as hypermutated (Fig. 1).

To assess the basis for the considerably different mutation rates, we evaluated MSI[7] and mutations in the DNA mismatch-repair pathway[8–10] genes *MLH1*, *MLH3*, *MSH2*, *MSH3*, *MSH6* and *PMS2*. Among the 30 hypermutated tumours with a complete data set, 23 (77%) had high levels of MSI (MSI-H). Included in this group were 19 tumours with *MLH1* methylation, 17 of which had CIMP. By comparison, the remaining seven hypermutated tumours, including the six with the highest mutation rates, lacked MSI-H, CIMP or MLH1 methylation but usually had somatic mutations in one or more mismatch-repair genes or *POLE* aberrations seen rarely in the non-hypermutated tumours (Fig. 1).

Gene mutations

Overall, we identified 32 somatic recurrently mutated genes (defined by MutSig[11] and manual curation) in the hypermutated and non-hypermutated cancers (Fig. 1b). After removal of non-expressed genes, there were 15 and 17 in the hypermutated and non-hypermutated cancers, respectively (Fig. 1b; for a complete list see Supplementary Table 3). Among the non-hypermutated tumours, the eight most frequently mutated genes were *APC*, *TP53*, *KRAS*, *PIK3CA*, *FBXW7*, *SMAD4*, *TCF7L2* and *NRAS*. As expected, the mutated *KRAS* and *NRAS* genes usually had oncogenic codon 12 and 13 or codon 61 mutations, whereas the remaining genes had inactivating mutations. *CTNNB1*, *SMAD2*, *FAM123B* (also known as *WTX*) and *SOX9* were also mutated frequently. *FAM123B* is an X-linked negative regulator of WNT signalling[12], and virtually all of its mutations were loss of function. Mutations in *SOX9*, a gene important for cell differentiation in the intestinal stem cell niche[13,14], have not been associated previously with human cancer, but all nine mutated alleles in the non-hypermutated CRCs were frameshift or nonsense mutations. Tumour-suppressor

*Lists of participants and their affiliations appear at the end of the paper.

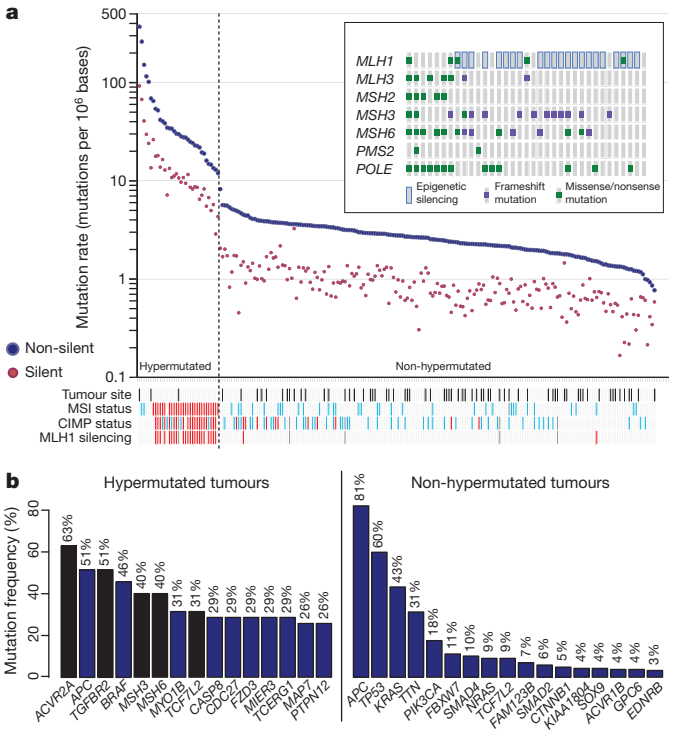


Figure 1 | Mutation frequencies in human CRC. a, Mutation frequencies in each of the tumour samples from 224 patients. Note a clear separation of hypermutated and non-hypermutated samples. Red, MSI high, CIMP high or MLH1 silenced; light blue, MSI low, or CIMP low; black, rectum; white, colon; grey, no data. Inset, mutations in mismatch-repair genes and *POLE* among the hypermutated samples. The order of the samples is the same as in the main graph. **b**, Significantly mutated genes in hypermutated and non-hypermutated tumours. Blue bars represent genes identified by the MutSig algorithm and black bars represent genes identified by manual examination of sequence data.

genes *ATM* and *ARID1A* also had a disproportionately high number of frameshift or nonsense mutations. *ARID1A* mutations have recently been reported in CRC and many other cancers[15,16].

In the hypermutated tumours, *ACVR2A*, *APC*, *TGFBR2*, *MSH3*, *MSH6*, *SLC9A9* and *TCF7L2* were frequent targets of mutation (Fig. 1b), along with mostly *BRAF(V600E)* mutations. However, two genes that were frequently mutated in the non-hypermutated cancers were significantly less frequently mutated in hypermutated tumours: *TP53* (60 versus 20%, $P < 0.0001$) and *APC* (81% versus 51%, $P = 0.0023$; both Fisher's exact test). Other genes, including *TGFBR2*, were mutated recurrently in the hypermutated cancers, but not in the non-hypermutated samples. These findings indicate that hypermutated and non-hypermutated tumours progress through different sequences of genetic events.

As expected, hypermutated tumours with MLH1 silencing and MSI-H showed additional differences in the mutational profile. When we specifically examined 28 genes with long mononucleotide repeats in their coding sequences, we found that the rate of frameshift mutation was 3.6-fold higher than the rate of such mutations in hypermutated tumours without MLH1 silencing and 50-fold higher than that in non-hypermethylated tumours (Supplementary Table 2).

Mutation rate and methylation patterns

As mentioned above, patients with colon and rectal tumours are managed differently[17], and epidemiology also highlights differences between the two[17]. An initial integrative analysis of MSI status, somatic copy-number alterations (SCNAs), CIMP status and gene-expression profiles of 132 colonic and 62 rectal tumours enabled us to examine possible biological differences between tumours in the two locations. Among the non-hypermutated tumours, however, the overall patterns of changes in copy number, CIMP, mRNA and miRNA were indistinguishable between colon and rectal carcinomas (Fig. 2). On the basis of this result, we merged the two for all subsequent analyses.

Unsupervised clustering of the promoter DNA methylation profiles of 236 colorectal tumours identified four subgroups (Supplementary Fig. 1 and Supplementary Methods). Two of the clusters contained tumours with elevated rates of methylation and were classified as CIMP high and CIMP low, as previously described[18]. The two non-CIMP clusters were predominantly from tumours that were non-hypermutated and derived from different anatomic locations. mRNA expression profiles separated the colorectal tumours into three distinct clusters (Supplementary Fig. 2). One significantly overlapped with CIMP-high tumours ($P = 3 \times 10^{-12}$) and was enriched with hypermutated tumours, and the other two clusters did not correspond with any group in the methylation data. Analysis of miRNA expression by unsupervised clustering (Supplementary Fig. 3) identified no clear distinctions between rectal cancers and non-hypermethylated colon cancers.

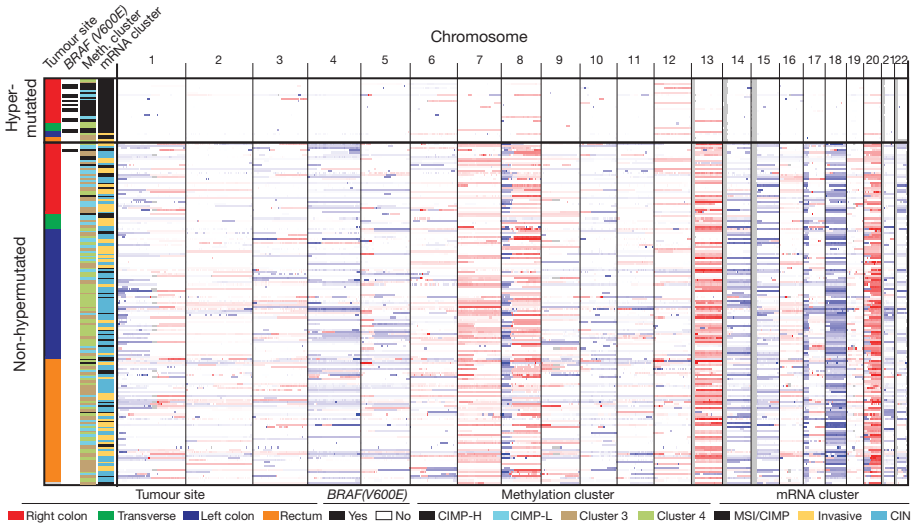


Figure 2 | Integrative analysis of genomic changes in 195 CRCs. Hypermutated tumours have near-diploid genomes and are highly enriched for hypermethylation, CIMP expression phenotype and *BRAF(V600E)* mutations. Non-hypermutated tumours originating from different sites are virtually indistinguishable from each other on the basis of their copy-number alteration patterns, DNA methylation or gene-expression patterns. Copy-number changes of the 22 autosomes are shown in shades of red for copy-number gains and shades of blue for copy-number losses.

Chromosomal and sub-chromosomal changes

In total, 257 tumours were profiled for SCNAs with Affymetrix SNP 6.0 arrays. Of these tumours, 97 were also analysed by low-depth-of-coverage (low-pass) whole-genome sequencing. As expected, the hypermutated tumours had far fewer SCNAs (Fig. 2). No difference was found between microsatellite-stable and -unstable hypermutated tumours (Supplementary Fig. 4). We used the GISTIC algorithm[19] to identify probable gene targets of focal alterations. There were several previously well-defined arm-level changes, including gains of 1q, 7p and q, 8p and q, 12q, 13q, 19q, and 20p and q (ref. 6). (Supplementary Fig. 4 and Supplementary Table 4). Significantly deleted chromosome arms were 18p and q (including *SMAD4*) in 66% of the tumours and 17p and q (including *TP53*) in 56%. Other significantly deleted chromosome arms were 1p, 4q, 5q, 8p, 14q, 15q, 20p and 22q.

We identified 28 recurrent deletion peaks (Supplementary Fig. 4 and Supplementary Table 4), including the genes *FHIT*, RBFOX1 and *WWOX* with large genomic footprints located in potentially fragile sites of the genome, in near-diploid hypermutated tumours. Other focal deletions involved tumour-suppressor genes such as *SMAD4*, *APC*, *PTEN* and *SMAD3*. A significant focal deletion of 10p25.2 spanned four genes, including *TCF7L2*, which was also frequently mutated in our data set. A gene fusion between adjacent genes *VTI1A* and *TCF7L2* through an interstitial deletion was found in 3% of CRCs and is required for survival of CRC cells bearing the translocation[4].

There were 17 regions of significant focal amplification (Supplementary Table 4). Some of these were superimposed on broad gains of chromosome arms, and included a peak at 13q12.13 near the peptidase-coding gene *USP12* and at ~500 kb distal to the CRC candidate oncogene *CDK8*; an adjacent peak at 13q12; a peak containing *KLF5* at 13q22.1; and a peak at 20q13.12 adjacent to *HNF4A*. Peaks on chromosome 8 included 8p12 (which contains the histone methyl-transferase-coding gene *WHSC1L1*, adjacent to *FGFR1*) and 8q24 (which contains *MYC*). An amplicon at 17q21.1, found in 4% of the tumours, contains seven genes, including the tyrosine kinase *ERBB2*. *ERBB2* amplifications have been described in colon, breast and gastro–oesophageal tumours, and breast and gastric cancers bearing these amplifications have been treated effectively with the anti-ERBB2 antibody trastuzumab[20–22].

One of the most common focal amplifications, found in 7% of the tumours, is the gain of a 100–150-kb region of the chromosome arm 11p15.5. It contains genes encoding insulin (*INS*), insulin-like growth factor 2 (*IGF2*) and tyrosine hydroxylase (*TH*), as well as *miR-483*, which is embedded within *IGF2* (Fig. 3a). We found elevated expression of *IGF2* and *miR-483* but not of *INS* and *TH* (Fig. 3b, c). Immediately adjacent to the amplified region is *ASCL2*, a transcription factor active in specifying intestinal stem-cell fate[23]. Although *ASCL2* has been implicated as a target of amplification in CRC[23–25], it was consistently outside the region of amplification and its expression was not correlated with copy-number changes. These observations suggest that *IGF2* and miR-483 are candidate functional targets of 11p15.5 amplification. *IGF2* overexpression through loss of imprinting has been implicated in the promotion of CRC[26, 27]. MiR-483 may also have a role in CRC pathogenesis[28].

A subset of tumours without *IGF2* amplification (15%) also had considerably higher levels of *IGF2* gene expression (as much as a 100-fold increase), an effect not attributable to methylation changes at the *IGF2* promoter. To assess the context of *IGF2* amplification/

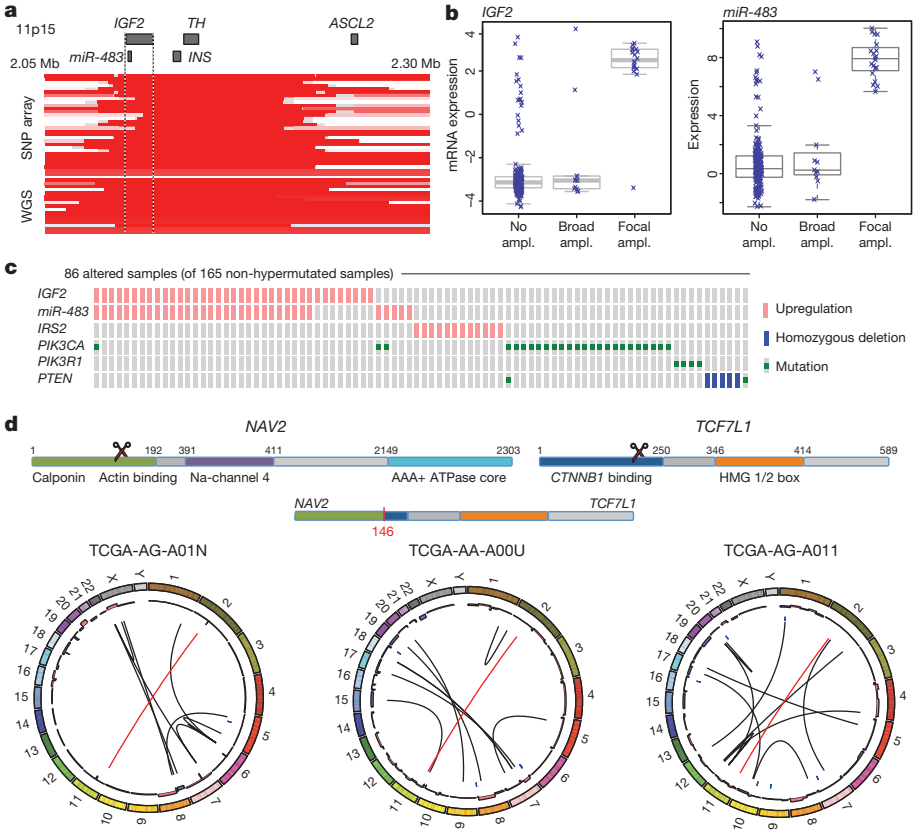


Figure 3 | Copy-number changes and structural aberrations in CRC. **a**, Focal amplification of 11p15.5. Segmented DNA copy-number data from single-nucleotide polymorphism (SNP) arrays and low-pass whole-genome sequencing (WGS) are shown. Each row represents a patient; amplified regions are shown in red. **b**, Correlation of expression levels with copy-number changes for *IGF2* and *miR-483*. **c**, *IGF2* amplification and overexpression are mutually exclusive of alterations in PI3K signalling-related genes. **d**, Recurrent *NAV2–TCF7L2* fusions. The structure of the two genes, locations of the breakpoints leading to the translocation and circular representations of all rearrangements in tumours with a fusion are shown. Red line lines represent the *NAV2–TCF7L2* fusions and black lines represent other rearrangements. The inner ring represents copy-number changes (blue denotes loss, pink denotes gain).

overexpression, we systematically searched for mutually exclusive genomic events using the MEMo method[29]. We found a pattern of near exclusivity (corrected $P < 0.01$) of *IGF2* overexpression with genomic events known to activate the PI3K pathway (mutations of *PIK3CA* and *PIK3R1* or deletion/mutation of *PTEN*; Fig. 3c and Supplementary Table 5). The *IRS2* gene, encoding a protein linking IGF1R (the receptor for IGF2) with PI3K, is on chromosome 13, which is frequently gained in CRC. The cases with the highest *IRS2* expression were mutually exclusive of the cases with *IGF2* overexpression ($P = 0.04$) and also lacked mutations in the PI3K pathway ($P = 0.0001$; Fig. 3c). These results strongly suggest that the IGF2–IGF1R–IRS2 axis signals to PI3K in CRC and imply that therapeutic targeting of the pathway could act to block PI3K activity in this subset of patients.

Translocations

To identify new chromosomal translocations, we performed low-pass, paired-end, whole-genome sequencing on 97 tumours with matched normal samples. In each case we achieved sequence coverage of ~3–4-fold and a corresponding physical coverage of 7.5–10-fold. Despite the low genome coverage, we detected 250 candidate interchromosomal translocation events (range, 0–10 per tumour). Among these events, 212 had one or both breakpoints in an intergenic region, whereas the remaining 38 juxtaposed coding regions of two genes in putative fusion events, of which 18 were predicted to code for in-frame events (Supplementary Table 6). We found three separate cases in which the first two exons of the *NAV2* gene on chromosome 11 are joined with the 3′ coding portion of *TCF7L1* on chromosome 2 (Supplementary Fig. 5). *TCF7L1* encodes TCF3, a member of the TCF/LEF class of transcription factors that heterodimerize with nuclear β-catenin to enable β-catenin-mediated transcriptional regulation. Intriguingly, in all three cases, the predicted structure of the NAV2–TCF7L1 fusion protein lacks the TCF3 β-catenin-binding domain. This translocation is similar to another recurrent translocation identified in CRC, a fusion in which the amino terminus of VTI1A is joined to TCF4, which is encoded by *TCF7L2*, a homologue of *TCF7L1* that is deleted or mutated in 12% of non-hypermutated tumours[4]. We also observed 21 cases of translocation involving *TTC28* located on chromosome 22 (Supplementary Table 6). In all cases the fusions predict inactivation of *TTC28*, which has been identified as a target of P53 and an inhibitor of tumour cell growth[30]. Eleven of the 19 (58%) gene–gene translocations were validated by obtaining PCR products or, in some cases, sequencing the junction fragments (Supplementary Fig. 5).

Altered pathways in CRC

Integrated analysis of mutations, copy number and mRNA expression changes in 195 tumours with complete data enriched our understanding of how some well-defined pathways are deregulated. We grouped samples by hypermutation status and identified recurrent alterations in the WNT, MAPK, PI3K, TGF-β and p53 pathways (Fig. 4, Supplementary Fig. 6 and Supplementary Table 1).

We found that the WNT signalling pathway was altered in 93% of all tumours, including biallelic inactivation of *APC* (Supplementary Table 7) or activating mutations of *CTNNB1* in ~80% of cases. There were also mutations in *SOX9* and mutations and deletions in *TCF7L2*, as well as the DKK family members and *AXIN2*, *FBXW7* (Supplementary Fig. 7), *ARID1A* and *FAM123B* (the latter is a negative regulator of WNT–β-catenin signalling[12] found mutated in Wilms' tumour[31]). A few mutations in *FAM123B* have previously been described in CRC[32]. SOX9 has been suggested to have a role in cancer, but no mutations have previously been described. The WNT receptor frizzled (*FZD10*) was overexpressed in ~17% of samples, in some instances at levels of 100× normal. Altogether, we found 16 different altered WNT pathway genes, confirming the importance of this pathway in CRC. Interestingly, many of these alterations were found in tumours that harbour *APC* mutations, suggesting that multiple lesions affecting the WNT signalling pathway confer selective advantage.

Genetic alterations in the PI3K and RAS–MAPK pathways are common in CRC. In addition to *IGF2* and *IRS2* overexpression, we found mutually exclusive mutations in *PIK3R1* and *PIK3CA* as well as deletions in *PTEN* in 2%, 15% and 4% of non-hypermutated tumours, respectively. We found that 55% of non-hypermutated tumours have alterations in *KRAS*, *NRAS* or *BRAF*, with a significant pattern of mutual exclusivity (Supplementary Fig. 6 and Supplementary Table 1). We also evaluated mutations in the erythroblastic leukemia viral oncogene homolog (ERBB) family of receptors because of the translational relevance of such mutations. Mutations or amplifications in

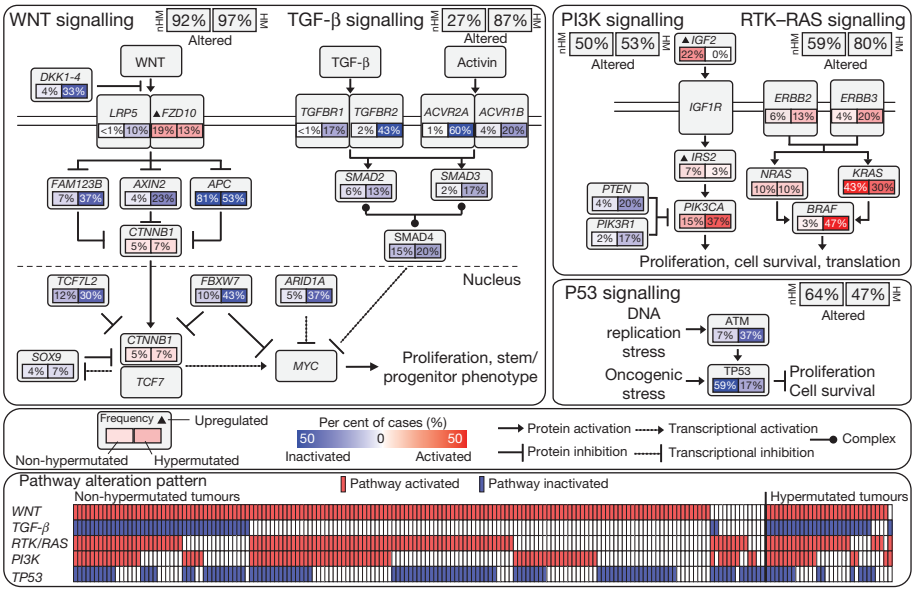


Figure 4 | Diversity and frequency of genetic changes leading to deregulation of signalling pathways in CRC. Non-hypermutated (nHM; $n = 165$) and hypermutated (HM; $n = 30$) samples with complete data were analysed separately. Alterations are defined by somatic mutations, homozygous deletions, high-level focal amplifications, and, in some cases, by significant up- or downregulation of gene expression (*IGF2*, *FZD10*, *SMAD4*). Alteration frequencies are expressed as a percentage of all cases. Red denotes activated genes and blue denotes inactivated genes. Bottom panel shows for each sample if at least one gene in each of the five pathways described in this figure is altered.

one of the four ERBB family genes are present in 22 out of 165 (13%) non-hypermutated and 16 out of 30 (53%) hypermutated cases. Some of the mutations are listed in the COSMIC database[33], suggesting a functional role. Intriguingly, recurrent *ERBB2(V842I)* and *ERBB3(V104M)* mutations were found in four and two non-hypermutated cases, respectively. Mutations and focal amplifications of *ERBB2* (Supplementary Fig. 6) should be evaluated as predictors of response to agents that target those receptors. We observed co-occurrence of alterations involving the RAS and PI3K pathways in one-third of tumours (Fig. 4; $P = 0.039$, Fisher's exact test). These results indicate that simultaneous inhibition of the RAS and PI3K pathways may be required to achieve therapeutic benefit.

The TGF-β signalling pathway is known to be deregulated in CRC and other cancers[34]. We found genomic alterations in *TGFBR1*, *TGFBR2*, *ACVR2A*, *ACVR1B*, *SMAD2*, *SMAD3* and *SMAD4* in 27% of the non-hypermutated and 87% of the hypermutated tumours. We also evaluated the p53 pathway, finding alterations in *TP53* in 59% of non-hypermutated cases (mostly biallelic; Supplementary Table 8) and alterations in ATM, a kinase that phosphorylates and activates P53 after DNA damage, in 7%. Alterations in these two genes showed a trend towards mutual exclusivity ($P = 0.016$) (Fig. 4, Supplementary Fig. 6 and Supplementary Table 1).

We integrated copy number, gene expression, methylation and pathway data using the PARADIGM software platform[35]. The analysis showed a number of new characteristics of CRC (Fig. 5a). For example, despite the diversity in anatomical origin or mutation levels, nearly 100% of these tumours have changes in MYC transcriptional targets, both those promoted by and those inhibited by MYC. These findings are consistent with patterns deduced from genetic alterations (Fig. 4) and suggest an important role for *MYC* in CRC. The analysis also identified several gene networks altered across all tumour samples and those with differential alterations in hypermutated versus non-hypermutated samples (Supplementary Table 7, Supplementary Data on the Cancer Genome Atlas publication webpage).

Because most of the tumours used in this study were derived from a prospective collection, survival data are not available. However, the tumours can be classified as aggressive or non-aggressive on the basis of tumour stage, lymph node status, distant metastasis and vascular invasion at the time of surgery. We found numerous molecular signatures associated with tumour aggressiveness, a subset of which is shown in Fig. 5b. They include specific focal amplifications and deletions, and altered gene-expression levels, including those of *SCN5A* (ref. 36), a reported regulator of colon cancer invasion (see Supplementary Tables 10 and 11 for a full list). Association with tumour aggressiveness is also observed in altered expression of miRNAs and specific somatic mutations (*APC*, *TP53*, *PIK3CA*, *BRAF* and *FBXW7*; Supplementary Fig. 8b). Mutations in *FBXW7* (38 cases) and distant metastasis (32 cases) never co-occurred ($P = 0.0019$). Interestingly, a number of genomic regions have multiple molecular associations with tumour aggressiveness that manifest as clinically related genomic hotspots. Examples of this are the region 20q13.12, which includes a focal amplification and multiple genes correlating with tumour aggression, and the region 22q12.3, containing *APOL6* (ref. 37) (Supplementary Figures 8 and 9).

Discussion

This comprehensive integrative analysis of 224 colorectal tumour and normal pairs provides a number of insights into the biology of CRC and identifies potential therapeutic targets. To identify possible biological differences in colon and rectum tumours, we found, in the non-hypermutated tumours irrespective of their anatomical origin, the same type of copy number, expression profile, DNA methylation and miRNA changes. Over 94% had a mutation in one or more members of the WNT signalling pathway, predominantly in *APC*. However, there were some differences between tumours from the right colon and all other sites. Hypermethylation was more common in the right colon, and three-quarters of hypermutated samples came from the same site, although not all of them had MSI (Fig. 2). Why most of the hypermutated samples came from the right colon and why there are two classes of tumours at this site is not known. The origins of the colon from embryonic midgut and hindgut may provide an explanation. As the survival rate of patients with high MSI-related cancers is better and these cancers are hypermutated, mutation rate may be a better prognostic indicator.

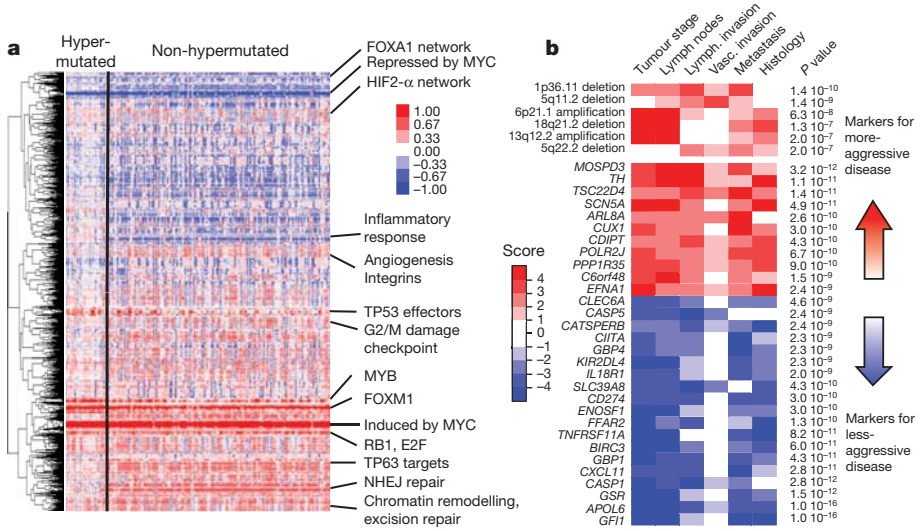


Figure 5 | Integrative analyses of multiple data sets. **a**, Clustering of genes and pathways affected in colon and rectum tumours deduced by PARADIGM analysis. Blue denotes under-expressed relative to normal and red denotes overexpressed relative to normal. Some of the pathways deduced by this method are shown on the right. NHEJ, non-homologous end joining. **b**, Gene-expression signatures and SCNAs associated with tumour aggression. Molecular signatures (rows) that show a statistically significant association with tumour aggressiveness according to selected clinical assays (columns) are shown in colour, with red indicating markers of tumour aggressiveness and blue indicating the markers of less-aggressive tumours. Significance is based on the combined *P* value from the weighted Fisher's method, corrected for multiple testing. Colour intensity and score is in accordance with the strength of an individual clinical–molecular association, and is proportional to $\log_{10}(P)$, where *P* is the *P* value for that association. To limit the vertical extent of the figure, gene-expression signatures are restricted to a combined *P* value of $P < 10^{-9}$ and SCNAs to $P < 10^{-7}$, and features are shown only if they are also significant in the subset of non-MSI-H samples (the analysis was performed separately on the full data as well as on the MSI-H and non-MSI-H subgroups).

Whole-exome sequencing and integrative analysis of genomic data provided further insights into the pathways that are dysregulated in CRC. We found that 93% of non-hypermutated and 97% of hypermutated cases had a deregulated WNT signalling pathway. New findings included recurrent mutations in *FAM123B*, *ARID1A* and *SOX9* and very high levels of overexpression of the WNT ligand receptor gene *FZD10*. To our knowledge, *SOX9* has not previously been described as frequently mutated in any human cancer. *SOX9* is transcriptionally repressed by WNT signalling, and the SOX9 protein has been shown to facilitate β-catenin degradation[38]. *ARID1A* is frequently mutated in gynaecological cancers and has been shown to suppress *MYC* transcription[39]. Activation of WNT signalling and inactivation of the TGF-β signalling pathway are known to result in activation of MYC. Our mutational and integrative analyses emphasize the critical role of MYC in CRC. We also compared our results with other large-scale analyses[6] and found many similarities and few differences in mutated genes (Supplementary Table 3).

Our integrated analysis revealed a diverse set of changes in TCF/LEF-encoding genes, suggesting additional roles for TCF/LEF factors in CRC beyond being passive partners for β-catenin.

Our data suggest a number of therapeutic approaches to CRC. Included are WNT-signalling inhibitors and small-molecule β-catenin inhibitors, which are showing initial promise[40–42]. We find that several proteins in the RTK–RAS and PI3K pathways, including IGF2, IGFR, ERBB2, ERBB3, MEK, AKT and MTOR could be targets for inhibition.

Our analyses show that non-hypermutated adenocarcinomas of the colon and rectum are not distinguishable at the genomic level. However, tumours from the right/ascending colon were more likely to be hypermethylated and to have elevated mutation rates than were other CRCs. As has been recognized previously, activation of the WNT signalling pathway and inactivation of the TGF-β signalling pathway, resulting in increased activity of MYC, are nearly ubiquitous events in CRC. Genomic aberrations frequently target the MAPK and PI3K pathways but less frequently target receptor tyrosine kinases. In conclusion, the data presented here provide a useful resource for understanding this deadly disease and identifying possibilities for treating it in a targeted way.

METHODS SUMMARY

Tumour and normal samples were processed by either of two biospecimen core resources, and aliquots of purified nucleic acids were shipped to the genome characterization and sequencing centres (Supplementary Methods). The biospecimen core resources provided sample sets in several different batches. To assess any batch effects we examined the mRNA expression, miRNA expression and DNA methylation data sets using a combination of cluster analysis, enhanced principal component analysis and analysis of variance (Supplementary Methods). Although some differences among batches were detected, we did not correct them computationally because the differences were generally modest and because some of them may reflect biological phenomena (Supplementary Methods).

We used Affymetrix SNP 6.0 microarrays to detect copy-number alterations. A subset of samples was subjected to low-pass (2–5×) whole-genome sequencing (Illumina HiSeq), in part for detection of SCNA and chromosomal translocations[43,44]. Gene-expression profiles were generated using Agilent microarrays and RNA-Seq. DNA methylation data were obtained using Illumina Infinium (HumanMethylation27) arrays. DNA sequencing of coding regions was performed by exome capture followed by sequencing on the SOLiD or Illumina HiSeq platforms. Details of the analytical methods used are described in Supplementary Methods.

All of the primary sequence files are deposited in dbGap and all other data are deposited at the Data Coordinating Center (DCC) for public access (http://cancergenome.nih.gov/). Data matrices and supporting data can be found at http://tcga-data.nci.nih.gov/docs/publications/coadread_2012/. The data can also be explored through the ISB Regulome Explorer (http://explorer.cancerregulome.org/), Next Generation Clustered Heat Maps (http://bioinformatics.mdanderson.org/main/TCGA/Supplements/NGCHM-CRC) and the cBio Cancer Genomics Portal (http://cbioportal.org). Descriptions of the data can be found at https://wiki.nci.nih.gov/x/j5dXAg and in Supplementary Methods.

Received 15 November 2011; accepted 22 May 2012.

1. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068 (2008).
2. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474,** 609–615 (2011).
3. Fearon, E. R. Molecular genetics of colorectal cancer. *Annu. Rev. Pathol.* **6,** 479–507 (2011).
4. Bass, A. J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent *VTI1A–TCF7L2* fusion. *Nature Genet.* **43,** 964–968 (2011).
5. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314,** 268–274 (2006).
6. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318,** 1108–1113 (2007).
7. Umar, A. *et al.* Revised Bethesda guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *J. Natl Cancer Inst.* **96,** 261–268 (2004).
8. Aaltonen, L. A. *et al.* Clues to the pathogenesis of familial colorectal cancer. 260, *Science* 812–816 (1993).
9. Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D. & Perucho, M. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature* **363,** 558–561 (1993).
10. Parsons, R. *et al.* Hypermutability and mismatch repair deficiency in RER^+ tumor cells. *Cell* **75,** 1227–1236 (1993).
11. Dooley, A. L. *et al.* Nuclear factor I/B is an oncogene in small cell lung cancer. *Genes Dev.* **25,** 1470–1475 (2011).
12. Major, M. B. *et al.* Wilms tumor suppressor WTX negatively regulates WNT/β-catenin signaling. *Science* **316,** 1043–1046 (2007).
13. Mori-Akiyama, Y. *et al.* SOX9 is required for the differentiation of paneth cells in the intestinal epithelium. *Gastroenterology* **133,** 539–546 (2007).
14. Bastide, P. *et al.* Sox9 regulates cell proliferation and is required for Paneth cell differentiation in the intestinal epithelium. *J. Cell Biol.* **178,** 635–648 (2007).
15. Jones, S. *et al.* Somatic mutations in the chromatin remodeling gene *ARID1A* occur in several tumor types. *Hum. Mutat.* **33,** 100–103 (2012).
16. Wilson, B. G. & Roberts, C. W. *SWI/SNF* nucleosome remodellers and cancer. *Nat. Rev. Cancer* **11,** 481–492 (2011).
17. Minsky, B. D. Unique considerations in the patient with rectal cancer. *Semin. Oncol.* **38,** 542–551 (2011).
18. Hinoue, T. *et al.* Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res.* **22,** 271–282 (2012).
19. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. USA* **104,** 20007–20012 (2007).
20. Camps, J. *et al.* Integrative genomics reveals mechanisms of copy number alterations responsible for transcriptional deregulation in colorectal cancer. *Genes Chromosom. Cancer* **48,** 1002–1017 (2009).
21. Varley, J. M., Swallow, J. E., Brammar, W. J., Whittaker, J. L. & Walker, R. A. Alterations to either *c-erb*B-2(neu) or *c-myc* proto-oncogenes in breast carcinomas correlate with poor short-term prognosis. *Oncogene* **1,** 423–430 (1987).
22. Yokota, J. *et al.* Amplification of *c-erb*B-2 oncogene in human adenocarcinomas *in vivo*. *Lancet* **327,** 765–767 (1986).
23. van der Flier, L. G. *et al.* Transcription factor achaete scute-like 2 controls intestinal stem cell fate. *Cell* **136,** 903–912 (2009).
24. Jubb, A. M., Hoeflich, K. P., Haverty, P. M., Wang, J. & Koeppen, H. Ascl2 and 11p15.5 amplification in colorectal cancer. *Gut* **60,** 1606–1607 (2011).
25. Stange, D. E. *et al.* Expression of an *ASCL2* related stem cell signature and *IGF2* in colorectal cancer liver metastases with 11p15.5 gain. *Gut* **59,** 1236–1244 (2010).
26. Cui, H. *et al.* Loss of *IGF2* imprinting: a potential marker of colorectal cancer risk. *Science* **299,** 1753–1755 (2003).
27. Nakagawa, H. *et al.* Loss of imprinting of the insulin-like growth factor II gene occurs by biallelic methylation in a core region of *H19*-associated CTCF-binding sites in colorectal cancer. *Proc. Natl Acad. Sci. USA* **98,** 591–596 (2001).
28. Veronese, A. *et al.* Oncogenic role of *miR-483-3p* at the *IGF2/483* locus. *Cancer Res.* **70,** 3140–3149 (2010).
29. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22,** 398–406 (2012).
30. Brady, C. A. *et al.* Distinct p53 transcriptional programs dictate acute DNA-damage responses and tumor suppression. *Cell* **145,** 571–583 (2011).
31. Rivera, M. N. *et al.* An X chromosome gene, *WTX*, is commonly inactivated in Wilms tumor. *Science* **315,** 642–645 (2007).
32. Scheel, S. K. *et al.* Mutations in the W*TX*-gene are found in some high-grade microsatellite instable (MSI-H) colorectal cancers. *BMC Cancer* **10,** 413 (2010).
33. Forbes, S. A. *et al.* The catalogue of somatic mutations in cancer (COSMIC). *Curr. Protoc. Hum. Genet.* Ch. 10, Unit 10.11 (2008).
34. Massagué, J., Blain, S. W. & Lo, R. S. TGFβ signaling in growth control, cancer, and heritable disorders. *Cell* **103,** 295–309 (2000).
35. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26,** i237–i245 (2010).
36. House, C. D. *et al.* Voltage-gated Na^+ channel *SCN5A* is a key regulator of a gene transcriptional network that controls colon cancer invasion. *Cancer Res.* **70,** 6957–6967 (2010).
37. Liu, Z., Lu, H., Jiang, Z., Pastuszyn, A. & Hu, C. A. Apolipoprotein l6, a novel proapoptotic Bcl-2 homology 3-only protein, induces mitochondria-mediated apoptosis in cancer cells. *Mol. Cancer Res.* **3,** 21–31 (2005).

38. Topol, L., Chen, W., Song, H., Day, T. F. & Yang, Y. Sox9 inhibits Wnt signaling by promoting β-catenin phosphorylation in the nucleus. *J. Biol. Chem.* **284,** 3323–3333 (2009).
39. Nagl, N. G. Jr, Zweitzig, D. R., Thimmapaya, B., Beck, G. R. Jr & Moran, E. The *c-myc* gene is a direct target of mammalian SWI/SNF-related complexes during differentiation-associated cell cycle arrest. *Cancer Res.* **66,** 1289–1293 (2006).
40. Chen, B. *et al.* Small molecule-mediated disruption of Wnt-dependent signaling in tissue regeneration and cancer. *Nat. Chem. Biol.* **5,** 100–107 (2009).
41. Ewan, K. *et al.* A useful approach to identify novel small-molecule inhibitors of Wnt-dependent transcription. *Cancer Res.* **70,** 5963–5973 (2010).
42. Sack, U. *et al.* S100A4-induced cell motility and metastasis is restricted by the Wnt/β-catenin pathway inhibitor calcimycin in colon cancer cells. *Mol. Biol. Cell* **22,** 3344–3354 (2011).
43. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6,** 677–681 (2009).
44. Xi, R. *et al.* Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc. Natl Acad. Sci. USA* **108,** E1128–E1136 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the following grants from the National Institutes of Health: U24CA143799, U24CA143835, U24CA143840, U24CA143843, U24CA143845, U24CA143848, U24CA143858, U24CA143866, U24CA143867, U24CA143882, U24CA143883, U24CA144025, U54HG003067, U54HG003079 and U54HG003273.

Author Contributions The Cancer Genome Atlas research network contributed collectively to this study. Biospecimens were provided by the tissue source sites and processed by the Biospecimen Core Resource. Data generation and analyses were performed by the genome-sequencing centers, cancer genome-characterization centers and genome data analysis centers. All data were released through the Data Coordinating Center. Project activities were coordinated by the National Cancer Institute and National Human Genome Research Institute project teams. Project leaders were R.K. and D.A.W. Writing team, T.A., A.J.B., T.A.C., L.D., A.H., S.R.H., R.K., P.W.L., M.M., N.S., I.S., J.M.S., J.T., V.T. and D.A.W.; mutations, M.S.L., L.R.T., D.A.W. and G.G.; copy-number and structural aberrations, A.H.R., A.J.B., A.H. and P.-C.C.; DNA methylation, T.H.; expression, J.T.A.; miRNA, G.R., A.C.; pathways, C.J.C., L.D., T.G., S.N., J.D.R., C.S., N.S., J.M.S. and V.T.

Author Information dbGaP accession numbers have been provided in Supplementary Table 1. The authors declare no competing financial interests. Reprints and permissions information is available at www.nature.com/reprints. Readers are welcome to comment on the online version of this article at www.nature.com/nature.

Correspondence and requests for materials should be addressed to R.K. (rkucherlapati@partners.org).

Genome Sequencing Center Baylor College of Medicine Donna M. Muzny[1], Matthew N. Bainbridge[1], Kyle Chang[1], Huyen H. Dinh[1], Jennifer A. Drummond[1], Gerald Fowler[1], Christie L. Kovar[1], Lora R. Lewis[1], Margaret B. Morgan[1], Irene F. Newsham[1], Jeffrey G. Reid[1], Jireh Santibanez[1], Eve Shinbrot[1], Lisa R. Trevino[1], Yuan-Qing Wu[1], Min Wang[1], Preethi Gunaratne[1,2], Lawrence A. Donehower[1,3], Chad J. Creighton[1,3], David A. Wheeler[1], Richard A. Gibbs[1]; **Genome Sequencing Center Broad Institute** Michael S. Lawrence[4], Douglas Voet[4], Rui Jing[4], Kristian Cibulskis[5], Andrey Sivachenko[3], Petar Stojanov[4], Aaron McKenna[4], Eric S. Lander[4,6,7], Stacey Gabriel[8], Gad Getz[4]; **Genome Sequencing Center Washington University in St Louis** Li Ding[9,10], Robert S. Fulton[9], Daniel C. Koboldt[9], Todd Wylie[9], Jason Walker[9], David J. Dooling[9,10], Lucinda Fulton[9], Kim D. Delehaunty[9], Catrina C. Fronick[9], Ryan Demeter[9], Elaine R. Mardis[9–11], Richard K. Wilson[9–11]; **Genome Characterization Center BC Cancer Agency** Andy Chu[12], Hye-Jung E. Chun[12], Andrew J. Mungall[12], Erin Pleasance[12], A. Gordon Robertson[12], Dominik Stoll[12], Miruna Balasundaram[12], Inanc Birol[12], Yaron S. N. Butterfield[12], Eric Chuah[12], Robin J. N. Coope[12], Noreen Dhalla[12], Ranabir Guin[12], Carrie Hirst[12], Martin Hirst[12], Robert A. Holt[12], Darlene Lee[12], Haiyan I. Li[12], Michael Mayo[12], Richard A. Moore[12], Jacqueline E. Schein[12], Jared R. Slobodan[12], Angela Tam[12], Nina Thiessen[12], Richard Varhol[12], Thomas Zeng[12], Yongjun Zhao[12], Steven J. M. Jones[12], Marco A. Marra[12]; **Genome-Characterization Center Broad Institute** Adam J. Bass[4,13], Alex H. Ramos[4,13], Gordon Saksena[4], Andrew D. Cherniack[4], Stephen E. Schumacher[4,13], Barbara Tabak[4,13], Scott L. Carter[4,13], Nam H. Pho[4], Huy Nguyen[4], Robert C. Onofrio[4], Andrew Crenshaw[4], Kristin Ardlie[4], Rameen Beroukhim[4,13], Wendy Winckler[4], Gad Getz[4], Matthew Meyerson[4,13,14]; **Genome-Characterization Center Brigham and Women's Hospital and Harvard Medical School** Alexei Protopopov[15], Juinhua Zhang[15], Angela Hadjipanayis[16,17], Eunjung Lee[17,18], Ruibin Xi[18], Lixing Yang[18], Xiaojia Ren[15], Hailei Zhang[15], Narayanan Sathiamoorthy[19], Sachet Shukla[15], Peng-Chieh Chen[16,17], Psalm Haseley[17,18], Yonghong Xiao[15], Semin Lee[18], Jonathan Seidman[16], Lynda Chin[4,15,20], Peter J. Park[17–19], Raju Kucherlapati[16,17]; **Genome-Characterization Center University of North Carolina, Chapel Hill** J. Todd Auman[21,22], Katherine A. Hoadley[23–25], Ying Du[25], Matthew D. Wilkerson[25], Yan Shi[25], Christina Liquori[25], Shaowu Meng[25], Ling Li[25], Yidi J. Turman[25], Michael D. Topal[24,25], Donghui Tan[26], Scot Waring[25], Elizabeth Buda[25], Jesse Walsh[25], Corbin D. Jones[27], Piotr A. Mieczkowski[23], Darshan Singh[28], Junyuan Wu[25], Anisha Gulabani[25], Peter Dolina[25], Tom Bodenheimer[25], Alan P. Hoyle[25], Janae V. Simons[25], Matthew Soloway[25], Lisle E. Mose[24], Stuart R. Jefferys[24], Saianand Balu[25], Brian D. O'Connor[25], Jan F. Prins[28], Derek Y. Chiang[23,25], D. Neil Hayes[25,29], Charles M. Perou[23–25]; **Genome-Characterization Centers University of Southern California and Johns Hopkins University** Toshinori Hinoue[30], Daniel J. Weisenberger[30], Dennis T. Maglinte[30], Fei Pan[30], Benjamin P. Berman[30], David J. Van Den Berg[30], Hui Shen[30], Timothy Triche Jr[30], Stephen B. Baylin[31], Peter W. Laird[30]; **Genome Data Analysis Center Broad Institute** Gad Getz[4], Michael Noble[4], Doug Voet[4], Gordon Saksena[4], Nils Gehlenborg[4,18], Daniel DiCara[4], Juinhua Zhang[4,15], Hailei Zhang[4,15], Chang-Jiun Wu[4,15], Spring Yingchun Liu[4,15], Sachet Shukla[4,15], Michael S. Lawrence[4], Lihua Zhou[4], Andrey Sivachenko[4], Pei Lin[4], Petar Stojanov[4], Rui Jing[4], Richard W. Park[18], Marc-Danie Nazaire[4], Jim Robinson[4], Helga Thorvaldsdottir[4], Jill Mesirov[4], Peter J. Park[17–19], Lynda Chin[4,15,20]; **Genome Data Analysis Center Institute for Systems Biology** Vesteinn Thorsson[32], Sheila M. Reynolds[32], Brady Bernard[32], Richard Kreisberg[32], Jake Lin[32], Lisa Iype[32], Ryan Bressler[32], Timo Erkkilä[32], Madhumati Gundapuneni[32], Yuexin Liu[33], Adam Norberg[32], Tom Robinson[32], Da Yang[33], Wei Zhang[33], Ilya Shmulevich[32]; **Genome Data Analysis Center Memorial Sloan-Kettering Cancer Center** Jorma J. de Ronde[34,35], Nikolaus Schultz[34], Ethan Cerami[34], Giovanni Ciriello[34], Arthur P. Goldberg[34], Benjamin Gross[34], Anders Jacobsen[34], Jianjiong Gao[34], Bogumil Kaczkowski[34], Rileen Sinha[34], B. Arman Aksoy[34], Yevgeniy Antipin[34], Boris Reva[34], Ronglai Shen[36], Barry S. Taylor[34], Timothy A. Chan[37], Marc Ladanyi[38], Chris Sander[34]; **Genome Data Analysis Center University of Texas MD Anderson Cancer Center** Rehan Akbani[39], Nianxiang Zhang[39], Bradley M. Broom[39], Tod Casasent[39], Anna Unruh[39], Chris Wakefield[39], Stanley R. Hamilton[33], R. Craig Cason[33], Keith A. Baggerly[39], John N. Weinstein[39,40]; **Genome Data Analysis Centers, University of California, Santa Cruz and the Buck Institute** David Haussler[41,42], Christopher C. Benz[43], Joshua M. Stuart[41], Stephen C. Benz[41], J. Zachary Sanborn[41], Charles J. Vaske[41], Jingchun Zhu[41], Christopher Szeto[41], Gary K. Scott[43], Christina Yau[43], Sam Ng[41], Ted Goldstein[41], Kyle Ellrott41, Eric Collisson[44], Aaron E. Cozen[41], Daniel Zerbino[41], Christopher Wilks[41], Brian Craft[41], Paul Spellman[45]; **Biospecimen Core Resource International Genomics Consortium** Robert Penny[46], Troy Shelton[46], Martha Hatfield[46], Scott Morris[46], Peggy Yena[46], Candace Shelton[46], Mark Sherman[46], Joseph Paulauskis[46]; **Nationwide Children's Hospital Biospecimen Core Resource** Julie M. Gastier-Foster[47–49], Jay Bowen[47], Nilsa C. Ramirez[47,48], Aaron Black[47], Robert Pyatt[47,48], Lisa Wise[47], Peter White[47,49]; **Tissue source sites and disease working group** Monica Bertagnolli[50], Jen Brown[51], Timothy A. Chan[52], Gerald C. Chu[53], Christine Czerwinski[51], Fred Denstman[54], Rajiv Dhir[55], Arnulf Dörner[56], Charles S. Fuchs[57,58], Jose G. Guillem[59], Mary Iacocca[51], Hartmut Juhl[60], Andrew Kaufman[52], Bernard Kohl III[61], Xuan Van Le[61], Maria C. Mariano[62], Elizabeth N. Medina[62], Michael Meyers[63], Garrett M. Nash[59], Phillip B. Paty[59], Nicholas Petrelli[54], Brenda Rabeno[51], William G. Richards[64], David Solit[66], Pat Swanson[51], Larissa Temple[52], Joel E. Tepper[65], Richard Thorp[61], Efsevia Vakiani[62], Martin R. Weiser[59], Joseph E. Willis[67], Gary Witkin[51], Zhaoshi Zeng[59], Michael J. Zinner[63], Carsten Zornig[68]; **Data-Coordination Center** Mark A. Jensen[69], Robert Sfeir[69], Ari B. Kahn[69], Anna L. Chu[69], Prachi Kothiyal[69], Zhining Wang[69], Eric E. Snyder[69], Joan Pontius[69], Todd D. Pihl[69], Brenda Ayala[69], Mark Backus[69], Jessica Walton[69], Jon Whitmore[69], Julien Baboud[69], Dominique L. Berton[69], Matthew C. Nicholls[69], Deepak Srinivasan[69], Rohini Raman[69], Stanley Girshik[69], Peter A. Kigonya[69], Shelley Alonso[69], Rashmi N. Sanbhadti[69], Sean P. Barletta[69], John M. Greene[69], David A. Pot[69]; **Project Team National Cancer Institute** Kenna R. Mills Shaw[70], Laura A. L. Dillon[70], Ken Buetow[71], Tanja Davidsen[71], John A. Demchok[70], Greg Eley[72], Martin Ferguson[73], Peter Fielding[70], Carl Schaefer[71], Margi Sheth[70] and Liming Yang[70]; **Project Team National Human Genome Research Institute** Mark S. Guyer[74], Bradley A. Ozenberger[74], Jacqueline D. Palchik[74], Jane Peterson[74], Heidi J. Sofia[74] & Elizabeth Thomson[74].

[1]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. [2]Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204, USA. [3]Dan L. Duncan Cancer Center, Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. [4]The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02142, USA. [5]Medical Sequencing Analysis and Informatics, The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02142, USA. [6]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [7]Department of Systems Biology, Harvard University, Boston, Massachusetts 02115, USA. [8]Genetic Analysis Platform, The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02142, USA. [9]The Genome Institute, Washington University School of Medicine, St Louis, Missouri 63108 USA. [10]Department of Genetics, Washington University School of Medicine, St Louis, Missouri 63108, USA. [11]Siteman Cancer Center, Washington University School of Medicine, St Louis, Missouri 63108, USA. [12]Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia V5Z 1L3, Canada. [13]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. [14]Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA. [15]Belfer Institute for Applied Cancer Science, Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. [16]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [17]Division of Genetics, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [18]The Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, USA. [19]Informatics Program, Children's Hospital, Boston, Massachusetts 02115, USA. [20]Department of Dermatology, Harvard Medical School, Boston, Massachusetts 02115, USA. [21]Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [22]Institute for Pharmacogenetics and Individualized Therapy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [23]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [24]Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599,

USA. [25]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [26]Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [27]Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [28]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [29]Department of Internal Medicine, Division of Medical Oncology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. [30]University of Southern California Epigenome Center, University of Southern California, Los Angeles, California 90089 USA. [31]Cancer Biology Division, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University, Baltimore, Maryland 21231, USA. [32]Institute for Systems Biology, Seattle, Washington 98109, USA. [33]Division of Pathology and Laboratory Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [34]Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [35]Divisions of Experimental Therapy, Molecular Biology, Surgical Oncology, The Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands. [36]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [37]Human Oncology and Pathogenesis Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [38]Department of Pathology, Human Oncology and Pathogenesis Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [39]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [40]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. [41]Department of Biomolecular Engineering and Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA. [42]Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, California 95064, USA. [43]Buck Institute for Age Research, Novato, California 94945, USA. [44]Division of Hematology/Oncology, University of California San Francisco, San Francisco, California 94143, USA. [45]Oregon Health and Science University, Department of Molecular and Medical Genetics, Portland, Oregon 97239, USA. [46]International Genomics Consortium, Phoenix, Arizona 85004, USA. [47]Nationwide Children's Hospital Biospecimen Core Resource, The Research Institute at Nationwide Children's Hospital, Columbus, Ohio 43205, USA. [48]The Ohio State University College of Medicine, Department of Pathology, Columbus, Ohio 43205, USA. [49]The Ohio State University College of Medicine, Department of Pediatrics, Columbus, Ohio 43205, USA. [50]Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, Brookline, Massachusetts 02115, USA. [51]Department of Pathology, Christiana Care Health Services, Newark, Delaware 19718, USA. [52]Human Oncology and Pathogenesis Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [53]Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Brookline, Massachusetts 02115, USA. [54]Department of Surgery, Helen F. Graham Cancer Center at Christiana Care, Newark, Delaware 19718, USA. [55]Department of Pathology, University of Pittsburgh, Pittsburgh, Pennsylvania 15213, USA. [56]Klinik für Chirurgie, Krankenhaus Alten Eichen, 22527 Hamburg, Germany. [57]Department of Medical Oncology, Dana-Farber Cancer Institute, Brookline, Massachusetts 02115, USA. [58]Department of Medicine, Brigham and Women's Hospital, Brookline, Massachusetts 02115, USA. [59]Department of Surgery, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [60]Indivumed Inc., Kensington, Maryland 20895, USA. [61]ILSbio, LLC, Chestertown, Maryland 21620, USA. [62]Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [63]Department of Surgery, Brigham and Women's Hospital, Brookline, Massachusetts 02115, USA. [64]Tissue and Blood Repository, Brigham and Women's Hospital, Brookline, Massachusetts 02115, USA. [65]Dept of Radiation Oncology, University of North Carolina School of Medicine. Chapel Hill, North Carolina 27599, USA. [66]Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. [67]Department of Pathology, Case Medical Center, Cleveland, Ohio 44106, USA. [68]Chirugische Klinik, Israelitisches Krankenhaus, 22297 Hamburg, Germany. [69]SRA International, Fairfax, Virginia 22033, USA. [70]The Cancer Genome Atlas Program Office, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [71]Center for Biomedical Informatics and Information Technology (CBIIT), National Cancer Institute, National Institutes of Health, Rockville, Maryland 20852, USA. [72]Scimentis, LLC, Statham, Georgia 30666, USA. [73]MLF Consulting, Arlington, Massachusetts 02474, USA. [74]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.

LETTER

doi:10.1038/nature11256

High velocity dispersion in a rare grand-design spiral galaxy at redshift $z = 2.18$

David R. Law[1], Alice E. Shapley[2], Charles C. Steidel[3], Naveen A. Reddy[4], Charlotte R. Christensen[5] & Dawn K. Erb[6]

Although grand-design spiral galaxies are relatively common in the local Universe, only one has been spectroscopically confirmed[1] to lie at redshift $z > 2$ (HDFX 28; $z = 2.011$); and it may prove to be a major merger that simply resembles a spiral in projection. The rarity of spirals has been explained as a result of disks being dynamically 'hot' at $z > 2$ (refs 2–5), which may instead favour the formation of commonly observed clumpy structures[6–10]. Alternatively, current instrumentation may simply not be sensitive enough to detect spiral structures comparable to those in the modern Universe[11]. At $z < 2$, the velocity dispersion of disks decreases[12], and spiral galaxies are more numerous by $z \approx 1$ (refs 7, 13–15). Here we report observations of the grand-design spiral galaxy Q2343-BX442 at $z = 2.18$. Spectroscopy of ionized gas shows that the disk is dynamically hot, implying an uncertain origin for the spiral structure. The kinematics of the galaxy are consistent with a thick disk undergoing a minor merger, which can drive the formation of short-lived spiral structure[16–18]. A duty cycle of <100 Myr for such tidally induced spiral structure in a hot massive disk is consistent with its rarity.

Using infrared imaging data from the Hubble Space Telescope Wide-Field Camera 3 (HST/WFC3), tracing rest-frame ~5,000-Å stellar continuum emission (details in Supplementary Information), we found that Q2343-BX442 (hereafter BX442) is well resolved, with a total luminous radius, R, of ~8 kpc, prominent spiral arms, a central nucleus, and a faint companion located 11 kpc away in projection to the northeast. These morphological characteristics (see summary in Table 1) led us to tentatively identify BX442 as a late-type Sc grand-design spiral galaxy. Strikingly, BX442 is the only object to display regular spiral morphology in a sample of 306 galaxies with similar imaging[10] at roughly the same redshift. We used the Keck/OSIRIS spectrograph in concert with the laser-guide-star adaptive optics (LGSAO) system to obtain integral field spectroscopy of nebular Hα emission from ionized gas regions within BX442 at an angular resolution comparable to that of the HST imaging data (~2 kpc; details in Supplementary Information). As shown in Fig. 1, the ionized gas emission similarly traces the structure of the spiral arms, but does not exhibit a prominent central bulge.

Stellar population model fits to broadband photometry (see Supplementary Information) indicate that the total star-formation rate (SFR) of BX442 is $52^{+37}_{-21}\ M_{\odot}\ \mathrm{yr}^{-1}$, with a characteristic population age of $1{,}100^{+1{,}000}_{-500}$ Myr and visual extinction $E(B-V) = 0.3 \pm 0.06$. BX442 is therefore drawn from the high end of the stellar mass function of $z \approx 2$ star-forming galaxies, with correspondingly higher than average size, dust extinction, stellar population age and SFR. This star formation is concentrated in the spiral arms, suggesting that their stellar populations may be significantly younger than those of the nucleus and inter-arm regions. Extrapolating from the Hα line-flux map using the Kennicutt relation[19,20], we estimate that the mean SFR surface density in the arms is $\Sigma_{\mathrm{SFR}} = 0.4\ M_{\odot}\ \mathrm{yr}^{-1}\ \mathrm{kpc}^{-2}$, peaking at $\sim 1\ M_{\odot}\ \mathrm{yr}^{-1}\ \mathrm{kpc}^{-2}$ in a bright clump located in the northern arm. Although these values are modest compared with characteristic values of Σ_{SFR} (~1–10 $M_{\odot}\ \mathrm{yr}^{-1}\ \mathrm{kpc}^{-2}$) that have been observed in detailed studies of the star-forming clumps of $z \approx 2$ galaxies[21], they are nonetheless ~30 times greater than typical for local spiral galaxies, and are similar to the values observed in local circumnuclear starbursts[19]. As indicated by rest-frame ultraviolet spectroscopy obtained with the Keck/LRIS spectrograph (details in Supplementary Information), the high Σ_{SFR} of BX442 drives outflows of gas into the surrounding intergalactic medium with speeds of up to a few hundred kilometres

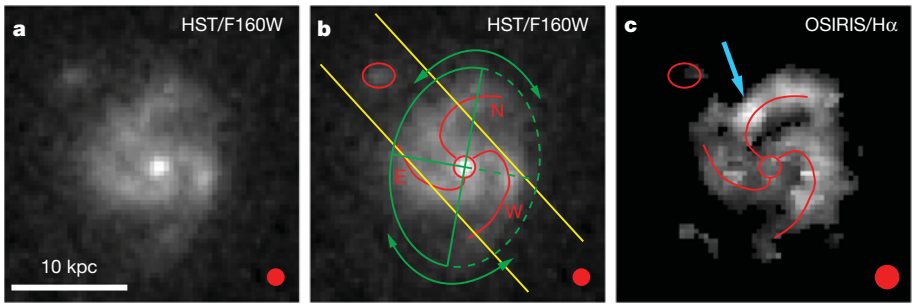


Figure 1 | Broadband and spectral emission-line morphology of BX442. **a**, **b**, HST/WFC3 F160W broadband morphology. In **b**, red lines show the locations of the northern (N), western (W) and eastern (E) spiral arms, core, and nearby satellite companion; green lines indicate the orientation of the best-fit inclined disk model (solid/dashed green lines represent opposite sides from the midplane); yellow lines represent the orientation of the long slit for previous Keck/NIRSPEC spectroscopy. The locations of these overlaid lines are defined visually; they are intended simply to guide the eye. (Alignment of individual images is discussed in Supplementary Information section 1.4.) **c**, Keck/OSIRIS Hα emission-line flux map, overlaid with the red lines from **b**. Blue arrow shows the location of a bright star-forming clump in the northern arm. A visual rejection criterion roughly correspondingly to a requirement for a signal-to-noise ratio of >3 (details in Supplementary Information) was used to mask low-flux pixels. The field of view in each panel (oriented with north up and east to the left) is 3 × 3 arcsec, corresponding to 25.3 × 25.3 kpc at the redshift of BX442. In each panel, the red dot shows the full-width at half-maximum (FWHM) of the observational point-spread function.

[1]Dunlap Institute for Astronomy & Astrophysics, University of Toronto, 50 St George Street, Toronto M5S 3H4, Ontario, Canada. [2]Department of Physics and Astronomy, University of California, Los Angeles, California 90095, USA. [3]California Institute of Technology, MS 249-17, Pasadena, California 91125, USA. [4]Department of Physics and Astronomy, University of California, Riverside, California 92521, USA. [5]Steward Observatory, 933 North Cherry Ave, Tucson, Arizona 85721, USA. [6]Department of Physics, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53211, USA.

per second, although there is no indication that the highest-Σ_{SFR} regions are the specific launching sites for galactic-scale outflows (see Supplementary Fig. 5), as recently proposed for a similar sample of galaxies[21].

Fitting a Gaussian profile to the Hα emission line at each location across the galaxy, we determined that the velocity profile of BX442 (Fig. 2) is consistent with the rotating disk hypothesis, and exhibits a smooth gradient of $\pm 150\,\mathrm{km\,s^{-1}}$ along the morphological major axis, with flux-weighted mean velocity dispersion $\sigma_m = 66 \pm 6\,\mathrm{km\,s^{-1}}$ (after correcting for the instrumental resolution). The faint companion detected in the HST image is spectroscopically confirmed to lie within $100\,\mathrm{km\,s^{-1}}$ of the systemic redshift of BX442, but does not follow the global rotational velocity field, and may therefore represent a merging dwarf galaxy with mass a few per cent of that of the primary (as determined from the rest-frame ~ 5,000-Å luminosity ratio). The velocity dispersion of the ionized gas in BX442 is highest in the spiral arms, and appears to peak at $\sigma = 113 \pm 14\,\mathrm{km\,s^{-1}}$ in a bright star-forming clump in the northern arm.

We constructed a three-dimensional inclined disk model (details in Supplementary Information) that accounts for observational effects such as the delivered point-spread function, and determined that BX442 is consistent (reduced $\chi^2 = 2.3$) with being a rotating disk inclined at $42 \pm 10^\circ$ to the line of sight, with an inclination-corrected circular velocity $v_c = 234^{+49}_{-29}\,\mathrm{km\,s^{-1}}$ at the outer edge of the disk ($R \approx 8$ kpc). As inferred from our best-fit model, the vertical velocity dispersion of the disk is $\sigma_z = 71\,\mathrm{km\,s^{-1}}$ ($v_c/\sigma_z \approx 3$), indicating that the system is geometrically thick, with a scale height $h_z = \sigma_z^2/(2\pi G\Sigma) = 0.7$ kpc, comparable to those of similarly massive systems studied in the literature[2,22,23]. (In the equation for scale height, G is the gravitational constant, and Σ is the mass surface density, here taken to be $3 \times 10^8\,M_\odot\,\mathrm{kpc^{-2}}$. The implied dynamical mass of BX442 is $M_{dyn} = 1.0^{+0.5}_{-0.2} \times 10^{11}\,M_\odot$ within a radius of $R = 8$ kpc, consistent with the sum of estimates of the gas ($M_{gas} = 2^{+2}_{-1} \times 10^{10}\,M_\odot$) and stellar ($M_* = 6^{+2}_{-1} \times 10^{10}\,M_\odot$) masses estimated from inversion of the Schmidt–Kennicutt law[19,20] and stellar population modelling, respectively.

Contrary to expectations[9,11,13], our observations of BX442 indicate both that dynamically hot $z \approx 2$ disk galaxies can form spiral structure, and that such structure can easily be detected with current-generation instruments such as HST/WFC3. Indeed, despite its high velocity dispersion, the surface density of BX442 is sufficiently high that the Toomre parameter[24] Q is ≤ 1 throughout most of the disk (details in Supplementary Information), suggesting that BX442 is susceptible to spontaneous formation of spiral structure. Galaxies with physical properties similar to those of BX442 are not remarkably uncommon at $z \approx 2$; large samples of galaxies with similar physical characteristics have been studied using high-angular-resolution imaging[25,26], integral-field spectroscopy[2,4,5], or both[5,27]. In particular, 27 galaxies in the recent morphology survey from which BX442 was drawn[10] have stellar masses within a factor of two of its mass, 10 of which also have similar half-light radii, star-formation rates, dust contents and stellar population ages. None of these other systems has clear spiral structure, indicating either that the triggering mechanism is relatively rare or that the duty cycle of the spiral pattern is extremely short.

Perhaps the most obvious distinction of BX442 is that it appears to be experiencing a close-passage minor merger, which numerical simulations and theoretical calculations suggest can be a natural means of producing grand-design spiral patterns in galactic disks[16,18,24], even for mass ratios as modest as a few per cent[17]. Indeed, many of the best known grand-design spiral galaxies in the nearby Universe (for example, M51, M81 and M101) are observed to have nearby companions, and small satellites such as the Sagittarius dwarf galaxy may even be partly responsible for producing spiral patterns in our own Milky Way galaxy[28]. We test the plausibility of the merger-induced hypothesis by comparing BX442 to a $z \approx 2$ model galaxy selected from a set of extremely high-resolution N-body smoothed particle hydrodynamic simulations[29] (details in Supplementary Information). Although the model disk spontaneously forms flocculent spiral structure in isolation, the lifetime of grand-design spiral patterns induced by the merging companion is generally less than half a rotation period (that is, $\leq$100 Myr, or $\Delta z \leq 0.08$ for BX442).

Such a mechanism may therefore naturally explain why visible spiral structure at $z \approx 2$ is so rare: not only must a galaxy be sufficiently massive to have stabilized the formation of an extended disk[30], but this disk must then be perturbed by a merging satellite sufficiently massive and properly oriented to excite an observable grand-design spiral

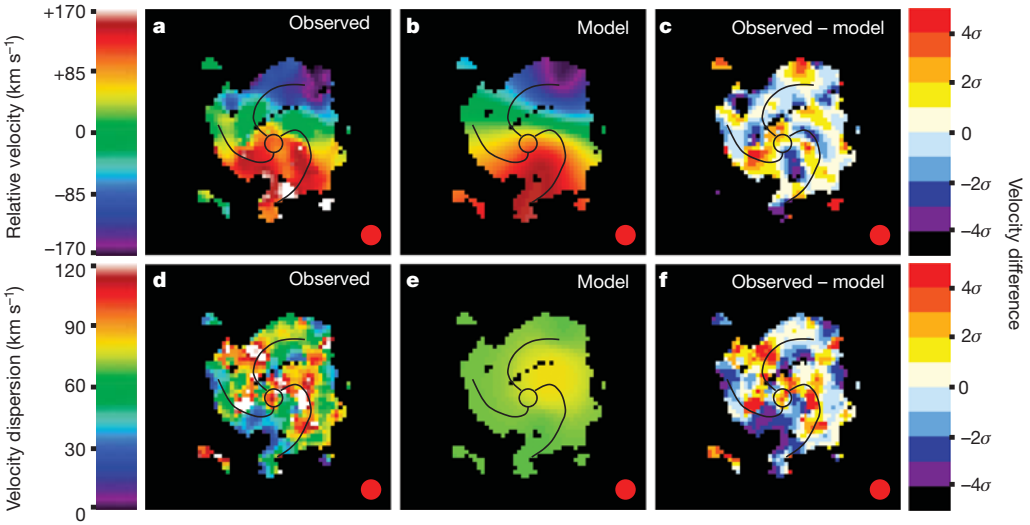


Figure 2 | Kinematic velocity and velocity-dispersion maps of BX442. **a**, **d**, Observed relative velocity (**a**) and velocity dispersion (**d**) (uncorrected for instrumental resolution), recovered from fitting Gaussian emission-line profiles to the Hα emission in each spatial pixel. The total integration time was 13 h, with a point-spread function (PSF) width of 0.25 arcsec (corresponding to ~2 kpc at the redshift of BX442). **b**, **e**, Best-fit inclined-disk models of the relative velocity (**b**) and velocity dispersion (**e**), after convolution with the observational PSF and Keck/OSIRIS spectral resolution. **c**, **f**, Residuals after subtraction of the best-fit models from the observed velocity (**c**) and velocity dispersion (**f**) fields. The residual values are given in units of the observational uncertainty: 1σ corresponds to $17\,\mathrm{km\,s^{-1}}$ for the line-of-sight velocity, and to $14\,\mathrm{km\,s^{-1}}$ for the line-of-sight velocity dispersion. Black lines indicating the spiral disk structure are overlaid from Fig. 1b; these lines indicate that the kinematic centre of BX442 is offset from the apparent nucleus of the continuum flux by ~2 kpc, owing in part to the uncertainty in image registration between the HST/WFC3 and Keck/OSIRIS data (see discussion in Supplementary Information). Red dots show the FWHM of the observational point-spread function.

Table 1 | Physical characteristics of BX442

Right ascension (J2000)	23 h 46 min 19.35 s
Declination (J2000)	+12° 48′ 00.0″
Redshift*	2.1765 ± 0.0001
Lookback time	10.7 Gyr
Stellar mass	$6^{+2}_{-1} \times 10^{10} M_{\odot}$
Gas mass	$2 \times 10^{10} M_{\odot}$
Age	$1{,}100^{+1{,}000}_{-500}$ Myr
SFR_{SED}	$52^{+37}_{-21}\ M_{\odot}\ yr^{-1}$
$SFR_{H\alpha}$	$45\ M_{\odot}\ yr^{-1}$
Bulge/total flux ratio†	10%
Inclination	42 ± 10°
Pitch angle‡	37 ± 6°
Spiral arm contrast§	1 AB $arcsec^{-2}$
Circular velocity	$234^{+49}_{-29}\ km\ s^{-1}$
Optical radius	8 kpc
Velocity dispersion	$71 \pm 1\ km\ s^{-1}$
Hubble type	Sc

See Supplementary Information for details. SFR_{SED} and $SFR_{H\alpha}$ are star-formation rates derived from stellar population modelling and inversion of the Schmidt–Kennicutt law, respectively.
* Derived from Hα nebular line emission and confirmed by multiple other emission and absorption line features.
† Decomposing the central nucleus from the surrounding disk using a model of the HST/WFC3 point-spread function indicates that the nuclear emission region contributes ~10% of the total rest-frame ~5,000-Å continuum flux, and has Sersic radial profile index $n \geq 4$ and a half-light radius $r \leq 1.5$ kpc, consistent with galactic bulges in nearby disk galaxies.
‡ Fourier phase-profile analysis of the spiral arms indicates substantial power in the $m = 2$ and $m = 3$ symmetry modes, corresponding to a three-armed spiral pattern (in which one arm is foreshortened by the inclination to the line of sight) with opening pitch angle $\alpha = 37 \pm 6°$.
§ Spiral arm/interarm surface brightness differential, in AB magnitudes per square arcsecond.

pattern. Further, this spiral must be observed in the narrow window of time for which its strength is at a maximum, and must be oriented sufficiently close to face-on that the pattern is recognizable.

Received 2 January; accepted 22 May 2012.

1. Dawson, S. *et al.* Optical and near-infrared spectroscopy of a high-redshift hard X-ray-emitting spiral galaxy. *Astron. J.* **125,** 1236–1246 (2003).
2. Genzel, R. *et al.* The rapid formation of a large rotating disk galaxy three billion years after the Big Bang. *Nature* **442,** 786–789 (2006).
3. Law, D. R. *et al.* Integral field spectroscopy of high-redshift star-forming galaxies with laser-guided adaptive optics: evidence for dispersion-dominated kinematics. *Astrophys. J.* **669,** 929–946 (2007).
4. Förster Schreiber, N. M. *et al.* The SINS survey: SINFONI integral field spectroscopy of z ~ 2 star-forming galaxies. *Astrophys. J.* **706,** 1364–1428 (2009).
5. Law, D. R. *et al.* The kiloparsec-scale kinematics of high-redshift star-forming galaxies. *Astrophys. J.* **697,** 2057–2082 (2009).
6. Conselice, C. J., Blackburne, J. A. & Papovich, C. The luminosity, stellar mass, and number density evolution of field galaxies of known morphology from z = 0.5 to 3. *Astrophys. J.* **620,** 564–583 (2005).
7. Elmegreen, D. M., Elmegreen, B. G., Rubin, D. S. & Schaffer, M. A. Galaxy morphologies in the Hubble Ultra Deep Field: dominance of linear structures at the detection limit. *Astrophys. J.* **631,** 85–100 (2005).
8. Law, D. R. *et al.* The physical nature of rest-UV galaxy morphology during the peak epoch of galaxy formation. *Astrophys. J.* **656,** 1–26 (2007).
9. Bournaud, F. & Elmegreen, B. G. Unstable disks at high redshift: evidence for smooth accretion in galaxy formation. *Astrophys. J.* **694,** L158–L161 (2009).
10. Law, D. R. *et al.* An HST/WFC3-IR morphological survey of galaxies at z = 1.5 - 3.6: I. Survey description and morphological properties of star forming galaxies. *Astrophys. J.* **745,** 85–122 (2012).
11. Conselice, C. J. *et al.* The tumultuous formation of the Hubble sequence at z > 1 examined with HST/Wide-Field Camera-3 observations of the Hubble Ultra Deep Field. *Mon. Not. R. Astron. Soc.* **417,** 2770–2788 (2011).
12. Wright, S. A. *et al.* Dynamics of galactic disks and mergers at z ~ 1.6: spatially resolved spectroscopy with Keck Laser Guide Star Adaptive Optics. *Astrophys. J.* **699,** 421–440 (2009).
13. Abraham, R. G. & van den Bergh, S. The morphological evolution of galaxies. *Science* **293,** 1273–1278 (2001).
14. Labbé, I. *et al.* Large disklike galaxies at high redshift. *Astrophys. J.* **591,** L95–L98 (2003).
15. Elmegreen, B. G., Elmegreen, D. M., Fernandez, M. X. & Lemonias, J. J. Bulge and clump evolution in Hubble Ultra Deep Field clump clusters, chains and spiral galaxies. *Astrophys. J.* **692,** 12–31 (2009).
16. Bottema, R. Simulations of normal spiral galaxies. *Mon. Not. R. Astron. Soc.* **344,** 358–384 (2003).
17. Dubinksi, J., Gauthier, J.-R., Widrow, L. & Nickerson, S. Spiral and bar instabilities provoked by dark matter satellites. *Astron. Soc. Pacific Conf. Series* **396,** 321–324 (2008).
18. Dobbs, C. L., Theis, C., Pringle, J. E. & Bate, M. R. Simulations of the grand design galaxy M51: a case study for analyzing tidally induced spiral structure. *Mon. Not. R. Astron. Soc.* **403,** 625–645 (2010).
19. Kennicutt, R. C. Jr. The global Schmidt law in star-forming galaxies. *Astrophys. J.* **498,** 541–552 (1998).
20. Bigiel, F. *et al.* The star formation law in nearby galaxies on sub-kpc scales. *Astron. J.* **136,** 2846–2871 (2008).
21. Genzel, R. *et al.* The SINS survey of z ~ 2 galaxy kinematics: properties of the giant star-forming clumps. *Astrophys. J.* **733,** 101–130 (2011).
22. Elmegreen, B. G. & Elmegreen, D. M. Observations of thick disks in the Hubble Space Telescope Ultra Deep Field. *Astrophys. J.* **650,** 644–660 (2006).
23. Genzel, R. *et al.* From rings to bulges: evidence for rapid secular galaxy evolution at z ~ 2 from integral field spectroscopy in the SINS survey. *Astrophys. J.* **687,** 59–77 (2008).
24. Toomre, A. in *The Structure and Evolution of Normal Galaxies* (eds Fall, S. M. & Lynden-Bell, D.) 111 (Cambridge University Press, 1981).
25. Elmegreen, D. M., Elmegreen, B. G., Ravindranath, S. & Coe, D. A. Resolved galaxies in the Hubble Ultra Deep Field: star formation in disks at high redshift. *Astrophys. J.* **658,** 763–777 (2007).
26. Kriek, M. *et al.* The Hubble sequence beyond z = 2 for massive galaxies: contrasting large star-forming and compact quiescent galaxies. *Astrophys. J.* **705,** L71–L75 (2009).
27. Förster Schreiber, N. M. *et al.* Constraints on the assembly and dynamics of galaxies. I. Detailed rest-frame optical morphologies on kiloparsec scale of z ~ 2 star-forming galaxies. *Astrophys. J.* **731,** 65–99 (2011).
28. Purcell, C. W. *et al.* The Sagittarius impact as an architect of spirality and outer rings in the Milky Way. *Nature* **477,** 301–303 (2011).
29. Wadsley, J. W., Stadel, J. & Quinn, T. Gasoline: a flexible, parallel implementation of Tree SPH. *N. Astron.* **9,** 137–158 (2004).
30. Martig, M. & Bournaud, F. Formation of late-type spiral galaxies: gas return from stellar populations regulates disk destruction and bulge growth. *Astrophys. J.* **714,** L275–L279 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements D.R.L and C.C.S have been supported by grant GO-11694 from the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS 5-26555. A.E.S acknowledges support from the David and Lucile Packard Foundation. C.R.C acknowledges support from the US National Science Foundation through grant AST-1009452. D.R.L appreciates discussions with J. Taylor, R. Abraham, J. Dubinski, F. Governato and A. Brooks, and thanks M. Peeples for help in obtaining the Keck/OSIRIS data.

Author Contributions D.R.L. performed the morphological analysis of the Hubble Space Telescope data and wrote the main manuscript text. The Keck/OSIRIS data were obtained by D.R.L. and A.E.S., and analysed by D.R.L. with extensive input from A.E.S. and C.C.S.. N.A.R. provided the Keck/LRIS spectra, Spitzer/MIPS photometry and stellar population modelling code, C.R.C. contributed the hydrodynamic galaxy simulations, and D.K.E. provided the Keck/NIRSPEC spectra. All authors reviewed, discussed and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.R.L. (drlaw@di.utoronto.ca).

LETTER

doi:10.1038/nature11246

How atomic nuclei cluster

J.-P. Ebran[1], E. Khan[2], T. Nikšić[3] & D. Vretenar[3]

Nucleonic matter displays a quantum-liquid structure, but in some cases finite nuclei behave like molecules composed of clusters of protons and neutrons. Clustering is a recurrent feature in light nuclei, from beryllium to nickel[1–3]. Cluster structures are typically observed as excited states close to the corresponding decay threshold; the origin of this phenomenon lies in the effective nuclear interaction, but the detailed mechanism of clustering in nuclei has not yet been fully understood. Here we use the theoretical framework of energy-density functionals[4,5], encompassing both cluster and quantum liquid-drop aspects of nuclei, to show that conditions for cluster formation can in part be traced back to the depth of the confining nuclear potential. For the illustrative example of neon-20, we show that the depth of the potential determines the energy spacings between single-nucleon orbitals in deformed nuclei, the localization of the corresponding wavefunctions and, therefore, the degree of nucleonic density clustering. Relativistic functionals, in particular, are characterized by deep single-nucleon potentials. When compared to non-relativistic functionals that yield similar ground-state properties (binding energy, deformation, radii), they predict the occurrence of much more pronounced cluster structures. More generally, clustering is considered as a transitional phenomenon between crystalline and quantum-liquid phases of fermionic systems.

The occurrence of molecular states in atomic nuclei and the formation of clusters of nucleons were predicted in the 1930s (refs 1 and 2). Subsequently, the description of nuclear dynamics came to be based predominantly on the concept of independent nucleons in a mean-field potential, but a renewed interest in clustering phenomena in the 1960s led to the development of theoretical methods dedicated to considering clusters[3]. Numerous experimental studies have revealed a wealth of data on clustering phenomena in light nuclei[3], and modern theoretical approaches use microscopic models that take single-nucleon degrees of freedom fully into account[6–8]. Clustering gives rise to nuclear molecules. For instance, in ^{12}C the second 0^+ state—the Hoyle state that has a key role in stellar nucleosynthesis—is predicted to display a structure composed of three α-particles[9,10]. The binding energy of the α-particle, formed from two protons and two neutrons, is much larger than that of other light nuclei. Cluster radioactivity[11], discovered in the 1980s, is another manifestation of clustering in atomic nuclei. Experimental signatures of clustering are usually indirect. Quasi-molecular resonances are probed by scattering one cluster on another, such as in the ^{12}C+^{12}C system[3,12], and cluster structures are also discernible in the break-up of nuclei. Evidence has been reported for the formation of clusters in ground and excited states of a number of α-conjugate nuclei[3]; that is, nuclei with an equal, even number of protons and neutrons, from ^{8}Be to ^{56}Ni.

The mechanism of cluster formation has not yet been fully understood. As shown in Ikeda diagrams[13], cluster structures are predicted to appear as excited states close to the corresponding decay threshold. However, the origin of cluster formation lies in the effective nuclear interaction, and signatures should also be present in the ground state[14–16]. Deformation has an important role because it removes the degeneracy of single-nucleon levels associated with spherical symmetry. At specific deformations the shell structure can restore degeneracies corresponding, for instance, to a 2:1 ratio of the large axis over the small axis of a quadrupole deformed system[3]. Consequently, the restored degeneracy of deformed shell closures facilitates the formation of clusters. However, this may be a rather qualitative explanation, because clustering phenomena cannot generally be explained by accidental degeneracies. Clustering is an essential feature of many-nucleon dynamics that coexists with the nuclear mean-field. Therefore, although in most cluster models the existence of such structures is assumed a priori and the corresponding effective interactions are adjusted to the binding energies and scattering phase shifts of these configurations, a fully microscopic understanding of cluster formation necessitates a more general description that encompasses both cluster and quantum liquid-drop aspects in light and heavier nuclei. It is well known that deformation and closeness to the cluster-emission threshold favour cluster formation. States close to the particle-emission threshold cannot be isolated from the environment of scattering states, so cluster states at the threshold belong to an open quantum system[17]. The aim of this work is to further explore the origin of clustering: to examine the conditions for cluster formation in ground states of finite nuclei, starting from a fully microscopic description based on the framework of energy-density functionals (EDFs).

At present, the only comprehensive approach to nuclear structure is based on the framework of EDFs. Nuclear EDFs enable a complete and accurate description of ground-state properties and collective excitations over the whole nuclide chart[4,5]. In practical implementations, nuclear EDFs are analogous to Kohn–Sham Density Functional Theory, the most widely used method for electronic-structure calculations in condensed-matter physics and quantum chemistry. In the nuclear case, the many-body dynamics is represented by independent nucleons moving in a local self-consistent mean-field potential that corresponds to the actual density and current distribution of a given nucleus. Both relativistic and non-relativistic realizations of EDFs are used in studies of nuclear matter and finite nuclei. A nuclear EDF is universal in the sense that, for a given inter-nucleon interaction, it has the same functional form for all systems. Using a small set of global parameters adjusted to empirical properties of homogeneous nuclear matter and data on finite nuclei, a universal functional provides a description of the structure of nuclei across the chart of nuclides.

A number of recent studies based on nuclear EDFs or the mean-field approach have analysed cluster structures in α-conjugate nuclei[14–16,18–20]. In Fig. 1 we display the self-consistent ground-state densities of ^{20}Ne, calculated with two widely used functionals that are representative of the two classes of nuclear EDFs: the non-relativistic Skyrme SLy4 (ref. 21), and the relativistic functional DD-ME2 (ref. 22). The equilibrium shape of ^{20}Ne is a prolate, axially symmetric quadrupole ellipsoid. Although they have not been specifically adjusted to this mass region, both functionals reproduce the empirical ground-state properties of this nucleus: the experimental binding energy, 160.6 MeV; the radius of the proton distribution, 2.90 fm (ref. 23); and the radius of the matter distribution, 2.85 fm (ref. 24), all with a typical accuracy to within roughly 1%. It is remarkable that, although these functionals predict similar values for the binding energy, charge and matter radii, and quadrupole deformation

[1]CEA/DAM/DIF, F-91297 Arpajon, France. [2]Institut de Physique Nucléaire, Université Paris-Sud, IN2P3-CNRS, F-91406 Orsay Cedex, France. [3]Physics Department, Faculty of Science, University of Zagreb, 10000 Zagreb, Croatia.

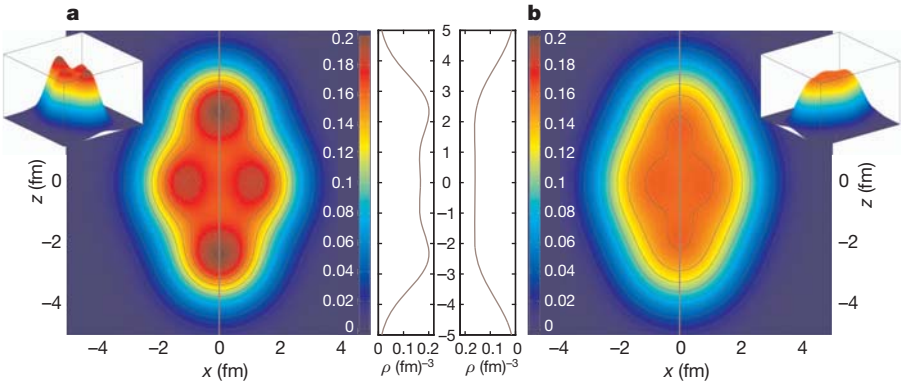


Figure 1 | Self-consistent ground-state densities of ^{20}Ne. Two nuclear energy-density functionals are used: **a**, DD-ME2 (ref. 22), and **b**, Skyrme SLy4 (refs 21 and 30). The densities (in units of fm^{-3}) are plotted in the $x-z$ plane of the intrinsic frame of reference that coincides with the principal axes of the nucleus, with z chosen as the symmetry axis. The inserts show the corresponding three-dimensional density plots and the density profiles (ρ) along the symmetry axis ($x = 0$).

of the equilibrium shape of ^{20}Ne, the corresponding single-nucleon densities are qualitatively very different. The density calculated with SLy4 displays a smooth behaviour characteristic of a Fermi liquid, with an extended surface region in which the density very gradually decreases from the central value of around 0.16 fm^{-3} (Fig. 1b). The relativistic functional DD-ME2, on the other hand, predicts an equilibrium density that is much more localized. The formation of cluster structures is clearly visible, with density spikes as large as roughly 0.2 fm^{-3}, and a much narrower surface region (Fig. 1a).

Understanding the difference in the equilibrium densities of ^{20}Ne calculated with SLy4 and DD-ME2 is a key to the mechanism of ground-state cluster formation in this mass region of α-conjugate deformed nuclei. The axially symmetric deformation of the nuclear mean-field removes the degeneracy of spherical single-nucleon levels, and nucleons paired by spin (up and down) occupy orbitals characterized by time-reversal degeneracy. For large deformations these levels can be labelled by a set of asymptotic Nilsson quantum numbers[25] and, because of the relatively weak Coulomb interactions in light nuclei, the localization of proton and neutron orbitals is similar in nuclei with equal numbers of protons and neutrons ($Z = N$ nuclei). In the specific case of ^{20}Ne, ten protons and ten neutrons occupy five deformed Nilsson levels, with the energy spacing between these levels proportional to the deformation of the single-nucleon potential. Figure 2 shows the partial single-nucleon densities that correspond to the highest occupied Nilsson orbital. Even without introducing a quantitative measure of localization, it is obvious that DD-ME2 predicts a much more localized density distribution (Fig. 2a). More-localized density distributions are also obtained for the other four occupied orbitals when calculated using DD-ME2.

Localization of densities that correspond to single-particle orbitals is a necessary precondition for the formation of clusters, and this effect can be traced back to the corresponding single-nucleon spectra. The comparison of spectra calculated with the two functionals shows that the one obtained with DD-ME2 is more spread out, and the more pronounced energy spacings between single-particle levels are also reflected in the more localized wavefunctions and partial densities. Starting from degenerate spherical single-particle levels, the splitting of the corresponding Nilsson deformed states is proportional to the deformation, and to the depth of the potential. Given that the two functionals predict almost identical equilibrium deformations and radii for ^{20}Ne, the different energy spacings in the single-nucleon spectra must have their origin in the difference in the corresponding potentials. In fact, the self-consistent mean-field potential of DD-ME2 is considerably deeper than that of SLy4. In the centre of the nucleus, the depth of the DD-ME2 single-neutron potential is −78.6 MeV, whereas the depth of the SLy4 potential is −69.5 MeV. The corresponding values of the single-proton potentials are −72.8 MeV for DD-ME2 and −64.6 MeV for SLy4. The effect of the potential depth on the localization of wavefunctions is shown schematically in Fig. 3a, where, as an approximation to nuclear potentials, we plot three harmonic-oscillator potentials with different depth values—30, 45 and 60 MeV—but the same radius, $R = 3$ fm. The radial wavefunctions of the corresponding p-states are shown in Fig. 3b. The oscillator length b determines the position of the maximum and the dispersion of the wavefunction[26]. The deeper the potential, the smaller the oscillator length (see the expression in the legend of Fig. 4), and the more localized the wavefunctions. In the classically forbidden region ($R > 3$ fm on Fig. 3), a smaller oscillator length leads to a more rapid

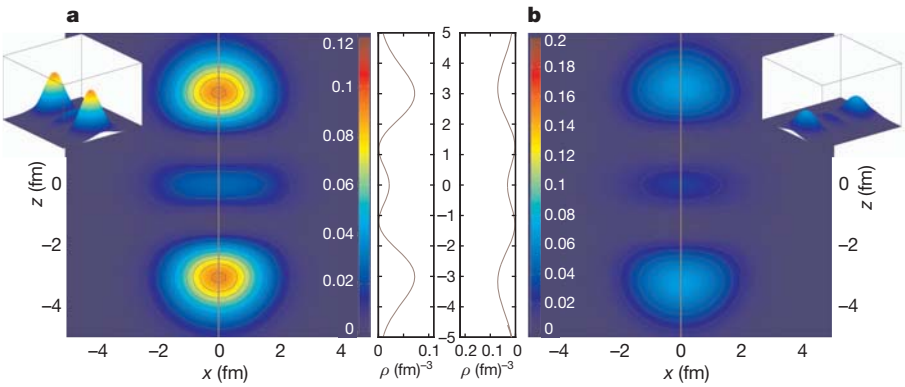


Figure 2 | Partial nucleon density distributions. Density distributions that correspond to the highest occupied level (2 protons spin up and down, and 2 neutrons spin up and down) in ^{20}Ne, having Nilsson quantum numbers $1/2^+[220]$, calculated using the nuclear energy-density functionals DD-ME2 (ref. 22) (**a**) and SLy4 (refs 21 and 30) (**b**).

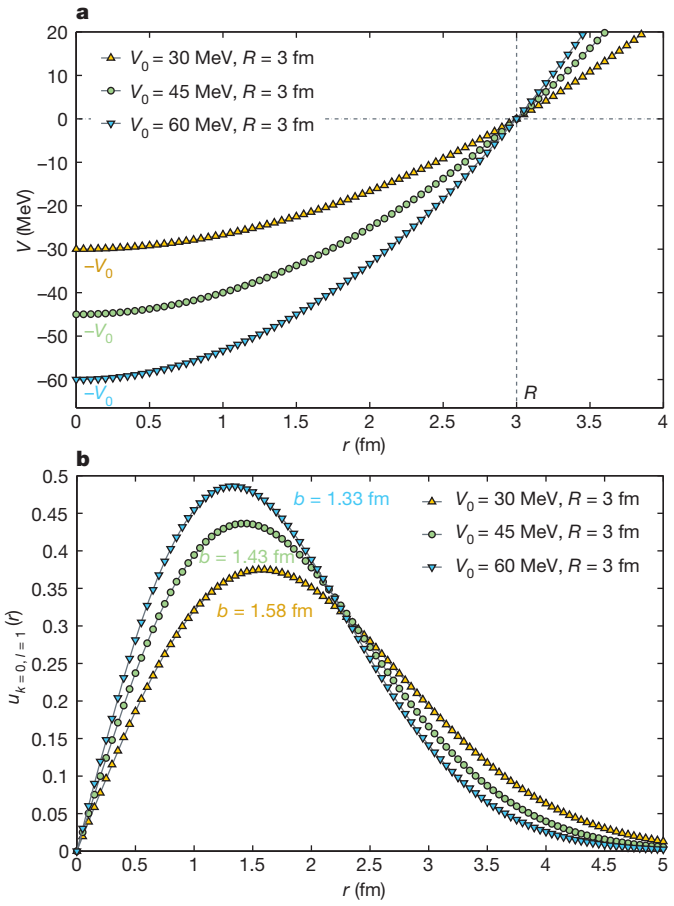


Figure 3 | Harmonic oscillators of different depths. **a**, Potentials plotted against the radial coordinate r for harmonic oscillators with the same radius, $R = 3$ fm, and depths (V_0) of 30, 45 and 60 MeV. **b**, The radial wavefunctions $U_{k,l}$ of the corresponding first p-state, where k is the radial quantum number and l the azimuthal one. The position of the maximum is determined by the oscillator length b.

exponential decay of the wavefunction, also favouring its localization. Hence a larger depth of the potential leads to a more pronounced localization of the wavefunction, in both the classically allowed and the forbidden regions, as shown in Fig. 3. In the present study we have verified, through a series of self-consistent mean-field calculations using a variety of non-relativistic and relativistic functionals for ^{20}Ne, ^{24}Mg, ^{28}Si and ^{32}S, that pronounced cluster structures in deformed equilibrium shapes indeed occur only for deep single-nucleon potentials.

The difference between the potential depths calculated with DD-ME2 and SLy4 is characteristic for relativistic versus non-relativistic self-consistent potentials. The depth of a relativistic potential is determined by the difference between two large fields: an attractive (negative) Lorentz scalar potential of magnitude around 400 MeV, and a repulsive Lorentz vector potential of roughly 320 MeV (plus the repulsive Coulomb potential for protons)[4,5]. In uniform matter these potentials are determined by the choice of the nuclear-matter equation of state; that is, by the density at which nucleonic matter saturates and by the binding energy per nucleon at saturation. The corresponding scalar and vector nucleon densities are related by a self-consistency condition[27] (in infinite matter the potentials are constant and proportional to the corresponding densities). Moreover, the sum of these potentials (about 700 MeV) determines the effective single-nucleon spin–orbit force in finite nuclear systems, which naturally manifests itself with the empirical strength. In a non-relativistic approach the spin–orbit potential is included in a purely phenomenological way, with the strength of the interaction adjusted to empirical energy spacings

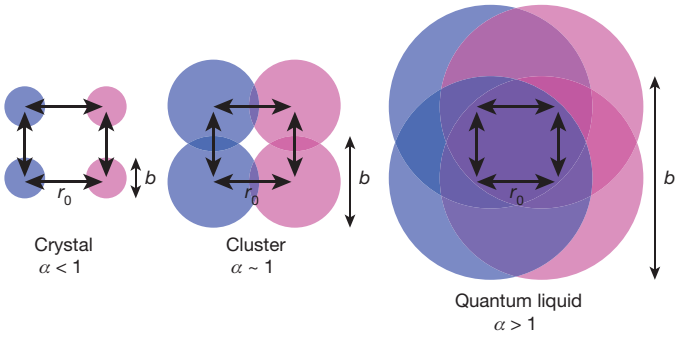


Figure 4 | Schematic illustration of the transition from a crystalline to a quantum liquid phase, including the cluster phase. The dimensionless parameter $\alpha = b/r_0$, where b is the dispersion of the fermion wavefunction and r_0 the typical inter-fermion distance, quantifies nuclear clustering. For a harmonic oscillator $\alpha = (\hbar R)^{1/2}\,(2mV_0)^{-1/4}\,r_0^{-1}$, where V_0 is the depth of the potential, R the radius of the system, m the mass of the nucleon and $\hbar$ Planck's constant/2π.

between spin–orbit partner states. Because the relativistic scalar and vector fields determine both the effective spin–orbit potential and the self-consistent single-nucleon mean-field, the latter is found to be deeper than the non-relativistic mean-field potentials for all relativistic functionals.

More generally, fermionic systems can exhibit a crystalline phase or, on the other extreme, a quantum liquid phase. The 'quantality' parameter has been thought[28] to show that nuclear matter displays a quantum-liquid structure. This concept can be generalized by considering nuclear clusters as transitional states between crystalline and quantum-liquid phases (Fig. 4). The dimensionless ratio $\alpha = b/r_0$, where b is the dispersion of the nucleon wavefunction and r_0 is the typical inter-nucleon distance (roughly 1.2 fm), is the natural parameter to quantify nuclear clustering, in analogy with similar considerations in condensed matter[29]. When α is greater than 1, nucleons are delocalized and the nucleus has a quantum-liquid structure. The transition to a cluster state occurs when α is about 1, so that nucleons become more localized and form a molecular structure (Fig. 4). In the present analysis we find that α is smaller than 1 for the relativistic functional, whereas it is greater than 1 for the non-relativistic functional. Moreover, from its definition in the case of a harmonic-oscillator potential (see legend of Fig. 4), α obviously increases with the number of nucleons (nuclear radius). Cluster states, therefore, are less likely to appear in heavier nuclei. The present discussion is also relevant for studies of the 'pasta' phase (located between the Wigner crystal and nuclear-matter phases) in the crust of neutron stars.

Received 6 March; accepted 17 May 2012.

1. Weizsäcker, C. F. V. Neuere Modellvorstellungen über den Bau der Atomkerne. *Naturwissenschaften* **26,** 209–217 (1938).
2. Wheeler, J. A. On the mathematical description of light nuclei by the method of resonating group structure. *Phys. Rev.* **52,** 1107–1122 (1937).
3. Von Oertzen, W. V., Freer, M. & Kanada-En'yo, Y. Nuclear clusters and nuclear molecules. *Phys. Rep.* **432,** 43–113 (2006).
4. Bender, M., Heenen, P.-H. & Reinhard, P.-G. Self-consistent mean-field models for nuclear structure. *Rev. Mod. Phys.* **75,** 121–180 (2003).
5. Vretenar, D., Afanasjev, A. V., Lalazissis, G. A. & Ring, P. Relativistic Hartree–Bogoliubov theory: static and dynamic aspects of exotic nuclear structure. *Phys. Rep.* **409,** 101–259 (2005).
6. Kanada-En'yo, Y. & Horiuchi, H. Structure of light unstable nuclei studied with antisymmetrized molecular dynamics. *Prog. Theor. Phys.* **142** (Suppl.), 205–263 (2001).
7. Feldmeier, H., Bieler, K. & Schnack, J. Fermionic molecular dynamics for ground states and collision of nuclei. *Nucl. Phys. A* **586,** 493–532 (1995).
8. Neff, T. & Feldmeier, H. Tensor correlations in the unitary correlation operator method. *Nucl. Phys. A* **713,** 311–371 (2003).
9. Tohsaki, A., Horiuchi, H., Schuck, P. & Röpke, G. Alpha cluster condensation in ^{12}C and ^{16}O. *Phys. Rev. Lett.* **87,** 192501 (2001).
10. Fynbo, H. O. U. *et al.* Revised rates for the stellar triple-α process from measurement of ^{12}C nuclear resonances. *Nature* **433,** 136–139 (2005).
11. Rose, H. J. & Jones, G. A. A new kind of natural radioactivity. *Nature* **307,** 245–247 (1984).
12. Greiner, W., Park, J. Y. & Scheid, W. *Nuclear Molecules* (World Scientific, 1995).

13. Ikeda K., Tagikawa N. & Horiuchi H.. The systematic structure-change into the molecule-like structures in the self-conjugate 4*n* nuclei. *Prog. Theor. Phys.* **464** (Suppl.), 464–475 (1968).
14. Arumugam, P., Sharma, B. K. & Patra, S. K. Relativistic mean field study of clustering in light nuclei. *Phys. Rev. C* **71,** 064308 (2005).
15. Maruhn, J. A. *et al.* α-Cluster structure and exotic states in a self-consistent model for light nuclei. *Phys. Rev. C* **74,** 044311 (2006).
16. Reinhard, P.-G., Maruhn, J. A., Umar, A. S. & Oberacker, V. E. Localization in light nuclei. *Phys. Rev. C* **83,** 034312 (2011).
17. Okolowicz, J., Ploszajczak, M. & Nazarewicz, W. On the origin of nuclear clustering. Preprint at ⟨http://arxiv.org/abs/1202.6290⟩ (2012).
18. Girod, M. & Grammaticos, B. Triaxial Hartree–Fock–Bogolyubov calculations with D1 effective interaction. *Phys. Rev. C* **27,** 2317–2339 (1983).
19. Ichikawa, T., Marhun, J. A., Itagaki, N. & Ohkubo, S. Linear chain structure of four-α clusters in ^{16}O. *Phys. Rev. Lett.* **107,** 112501 (2011).
20. Robledo, L. M. & Bertsch, G. F. Global systematics of octupole excitations in even–even nuclei. *Phys. Rev. C* **84,** 054302 (2011).
21. Chabanat, E., Bonche, P., Haensel, P., Meyer, J. & Schaeffer, R. A Skyrme parametrization from subnuclear to neutron star densities part II. Nuclei far from stabilities. *Nucl. Phys. A* **635,** 231–256 (1998).
22. Lalazissis, G. A., Nikšić, T., Vretenar, D. & Ring, P. New relativistic mean-field interaction with density-dependent meson-nucleons couplings. *Phys. Rev. C* **71,** 024312 (2005).
23. Fricke, G. *et al.* Behavior of the nuclear charge radii systematics in the *s-d* shell from muonic atom measurement. *Phys. Rev. C* **45,** 80–89 (1992).
24. Chulkov, L. *et al.* Interaction cross sections and matter radii of $A = 20$ isobars. *Nucl. Phys. A* **603,** 219–237 (1996).
25. Nilsson, S. G. Binding states of individual nucleons in strongly deformed nuclei *Mat. Fys. Medd. Dan. Vid. Selsk.* **29,** 1–69 (1955).
26. Cohen-Tannoudji, C., Diu, B. & Laloë, F. *Mécanique Quantique* (Hermann Ed., 1973).
27. Walecka, J. D. *Theoretical Nuclear and Subnuclear Physics* (Imperial College Press and World Scientific, 2004).
28. Mottelson, B. Elementary features of nuclear structure. In *Les Houches Session LXVI, Trends in Nuclear Physics, 100 Years Later* (eds Nifenecker, H., Blaizot, J.-P., Bertsch, G. F., Weise, W. & David, F.) 25–121 (North-Holland Elsevier, 1996).
29. Pines, D. & Nozières, P. *The theory of quantum liquids* (Benjamin, 1966).
30. Stoitsov, M. V., Dobaczewski, J., Nazarewicz, W. & Ring, P. Axially deformed solution of the Skyrme–Hartree–Fock–Bogolyubov equations using the transformed harmonic oscillator basis. The program HFBTHO (v1.66p). *Comput. Phys. Commun.* **167,** 43–63 (2005).

Acknowledgements This work was supported by the Institut Universitaire de France and by the Croatian Ministry of Science, Education and Sport—project 1191005-1010. The authors thank J. Margueron, M. Milin, T. Neff, N. Van Giai and P. Schuck for comments and suggestions.

Author Contributions Model calculations were done by J.-P.E., E.K., T.N. and D.V. The manuscript text was prepared by E.K. and D.V. with contributions from J.-P.E. and T.N. J.-P.E. and E.K. prepared the figures.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to E.K. (khan@ipno.in2p3.fr).

LETTER

doi:10.1038/nature11231

Terahertz-field-induced insulator-to-metal transition in vanadium dioxide metamaterial

Mengkun Liu[1]*, Harold Y. Hwang[2]*, Hu Tao[3], Andrew C. Strikwerda[1], Kebin Fan[4], George R. Keiser[1], Aaron J. Sternbach[1], Kevin G. West[5], Salinporn Kittiwatanakul[5], Jiwei Lu[5], Stuart A. Wolf[5,6], Fiorenzo G. Omenetto[3], Xin Zhang[4], Keith A. Nelson[2] & Richard D. Averitt[1]

Electron–electron interactions can render an otherwise conducting material insulating[1], with the insulator–metal phase transition in correlated-electron materials being the canonical macroscopic manifestation of the competition between charge-carrier itinerancy and localization. The transition can arise from underlying microscopic interactions among the charge, lattice, orbital and spin degrees of freedom, the complexity of which leads to multiple phase-transition pathways. For example, in many transition metal oxides, the insulator–metal transition has been achieved with external stimuli, including temperature, light, electric field, mechanical strain or magnetic field[2–7]. Vanadium dioxide is particularly intriguing because both the lattice and on-site Coulomb repulsion contribute to the insulator-to-metal transition at 340 K (ref. 8). Thus, although the precise microscopic origin of the phase transition remains elusive, vanadium dioxide serves as a testbed for correlated-electron phase-transition dynamics. Here we report the observation of an insulator–metal transition in vanadium dioxide induced by a terahertz electric field. This is achieved using metamaterial-enhanced picosecond, high-field terahertz pulses to reduce the Coulomb-induced potential barrier for carrier transport[9]. A nonlinear metamaterial response is observed through the phase transition, demonstrating that high-field terahertz pulses provide alternative pathways to induce collective electronic and structural rearrangements. The metamaterial resonators play a dual role, providing sub-wavelength field enhancement that locally drives the nonlinear response, and global sensitivity to the local changes, thereby enabling macroscopic observation of the dynamics[10,11]. This methodology provides a powerful platform to investigate low-energy dynamics in condensed matter and, further, demonstrates that integration of metamaterials with complex matter is a viable pathway to realize functional nonlinear electromagnetic composites.

Ultrafast spectroscopic techniques are important for investigating phase-transition dynamics, because they can be used to initiate changes, and provide sufficient time resolution to monitor excited states or metastable order parameters not accessible with time-integrated measurements[6,7,12,13]. Indeed, recent ultrafast pump–probe measurements on vanadium dioxide (VO_2) with near-infrared pulses revealed that excitation of electrons across the insulating Hubbard gap results in percolative metallicity on a picosecond timescale[12,13]. Direct-current electric fields of $\sim$100 kV cm^{-1}also induce the insulator-to-metal transition (IMT), but do not permit measurement of the field-induced transition dynamics[9,14]. From existing direct current (d.c.) measurements it has not been clear whether ultrafast electric fields could induce the IMT, or what the timescale would be.

Recent developments have enabled the generation of ultrafast terahertz (THz) pulses with field levels of 0.1–1 MV cm^{-1} (refs 15–17). Time-resolved THz pump–probe measurements have revealed dynamic electronic responses initiated by such high fields[15,18]. THz probe pulses are ideally suited to monitor the IMT, as they provide a direct measure of the conductivity in the GHz–THz frequency range. Figure 1a shows the temperature-dependent far-infrared conductivity (σ_1) of our 75-nm VO_2 film deposited on a sapphire substrate[19]. The results were obtained by fitting a Drude response to the transmission from a conventional THz time-domain spectroscopy (THz-TDS) measurement using low-field THz pulses[6,13]. The IMT occurs at 340 K, and in the high-temperature rutile phase the conductivity of $\sim$5,000 (Ω cm)$^{-1}$ is comparable to that of bulk single-crystal VO_2. Hysteresis, associated with the first-order structural transition, is also observed. Below 330 K, the conductivity is below our detection limit with conventional THz-TDS. For intense THz excitation, we deposited metamaterial structures that served as local resonant THz concentrators. Using gold split-ring resonators (SRRs), which are essentially sub-wavelength *LC* circuits, the THz electric field inside the SRR capacitive gap can be enhanced by more than an order of magnitude.

Figure 1b shows an optical image of the 200-nm-thick gold SRRs that we deposited onto a VO_2 film. The SRR lateral dimension is 76 μm, with a periodicity of 100 μm. For our experiments, the most important regions are the 1.5-μm SRR capacitive gaps that are oriented horizontally, where the vertically polarized THz field is enhanced. For these SRRs, the *LC* resonance, shown in Fig. 1c (300-K data) is at 0.41 THz, whereas the pure electric-dipole resonance (which would occur in the gap with just the two adjacent vertical gold segments forming a dipole antenna) is at $\sim$1 THz (ref. 20). In addition to THz field enhancement in the gaps, the SRR structures provide exquisite sensitivity for THz probing of small changes in the VO_2 thin-film transmission near the phase transition, where the conductivity is small. Although only a small fraction of the THz probe radiation directly irradiates the gaps, the resonant behaviour of the entire SRR array is affected profoundly by the in-gap VO_2 properties. Figure 1c shows the frequency-dependent SRR/VO_2 response as a function of temperature, measured using low-field THz-TDS. The SRR gap is gradually shorted as the VO_2 becomes metallic, leading to a transmission increase at the resonance frequency with increasing temperature. The inductive-capacitive resonance *LC* disappears at $\sim$350 K, corresponding to a film conductivity of $\sim$200 (Ω cm)$^{-1}$. From 300 to 340 K, there is a small but perceptible redshift of the *LC* resonance due to increasing permittivity of the VO_2 film associated with percolation of the metallic phase. A small hysteresis in the transmission is also observed (not shown), analogous to that shown in Fig. 1a and observed in previous experiments[21]. The higher-frequency dipolar resonance also shows a redshift, because the effective dipole length increases as the SRR gaps are shorted. As Fig. 1c shows, the SRRs provide enhanced conductivity sensitivity to changes in the material properties within the gaps. Thus,

[1]Department of Physics, Boston University, Boston, Massachusetts 02215, USA. [2]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139 USA. [3]Biomedical Engineering, Tufts University, Medford, Massachusetts 02155, USA. [4]Department of Mechanical Engineering, Boston University, Boston, Massachusetts 02215, USA. [5]Department of Materials Science and Engineering, University of Virginia, Charlottesville, Virginia 22904, USA. [6]Department of Physics, University of Virginia, Charlottesville, Virginia 22904, USA.
* These authors contributed equally to this work.

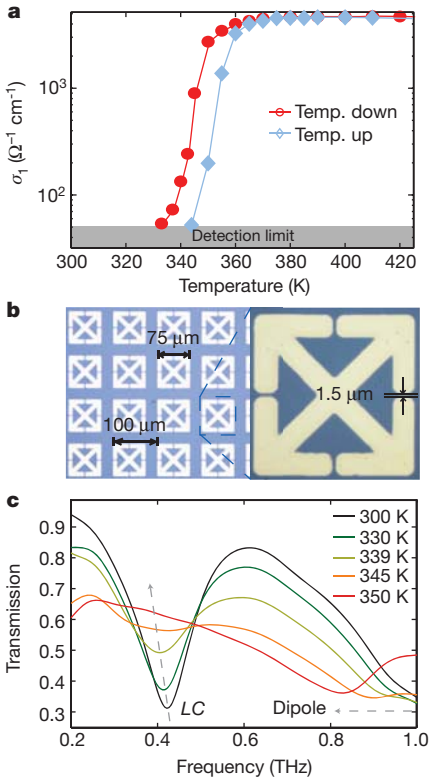


Figure 1 | Low-field THz characterization of 75-nm VO_2 thin film on sapphire with and without metamaterials. **a**, Temperature-dependent far-infrared conductivity (σ_1) of bare VO_2 thin film (75 nm thick) on sapphire substrate measured by THz-TDS. σ_1 was obtained by fitting the THz transmission to a standard Drude response. **b**, Optical image of metamaterial split-ring resonators (SRRs) deposited on VO_2/sapphire. The SRR gap is 1.5 μm. **c**, Temperature-dependent THz transmission spectra of SRRs on VO_2. *LC* and dipole resonances as described in text.

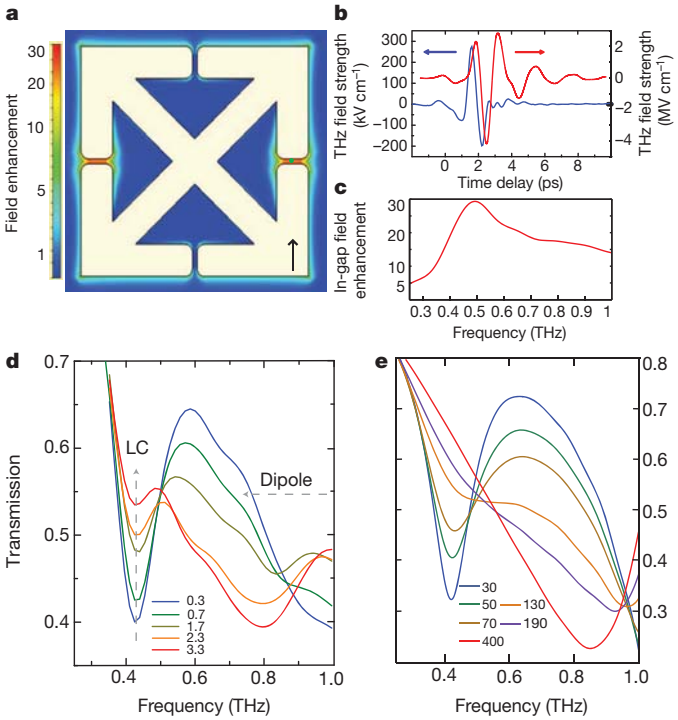


Figure 2 | Full-wave simulations of the electric field enhancement in the SRR and nonlinear THz transmission experiment. **a**, Resonant field enhancement as a function of position. **b**, Simulated time-dependent THz field (red) in the horizontal gaps using experimental data (blue) as the input. **c**, Frequency-dependent in-gap field enhancement obtained from the ratio of Fourier amplitudes of the simulated in-gap and measured incident fields in **b**. **d**, Experimental data showing field-dependent nonlinear transmission spectra of SRRs on VO_2 at 324 K, for in-gap fields ranging from 0.3 to 3.3 MV cm^{-1}. **e**, Full-wave simulations of SRR response for in-gap conductivities ranging from 30 to 400 $(\Omega\,\mathrm{cm})^{-1}$ (assuming σ_1 changes only in the gaps).

in the following nonlinear dynamics measurements, SRRs provide local excitation by field enhancement in the capacitive gaps and global sensitivity to the induced changes in the VO_2 within the gaps.

Full-wave electromagnetic simulations reveal the spatiotemporal features of the field enhancement in the SRRs (see Supplementary Information). Figure 2a shows that in the horizontally oriented gaps, there is a 27-fold field enhancement at the *LC* resonance frequency. Figure 2b shows the simulated time-dependent electric field (red curve) in the middle of the horizontal gaps, with the experimentally measured THz field (blue curve) used as input for the simulation. An incident peak field amplitude of ~300 kV cm^{-1} leads to a peak field of 4 MV cm^{-1} in the gap. Fourier transformation of the time-domain field profiles in Fig. 2b and calculation of the ratio of the spectral amplitudes yields the field enhancement as a function of frequency, as shown in Fig. 2c. The field enhancement is quite broadband as a result of the breadth and close proximity of the *LC* and electric-dipole modes. The magnitude of the calculated field enhancement is consistent with calculated and experimental results recently obtained on simpler antenna structures[20,22,23].

Figure 2d shows the experimentally measured nonlinear response of SRRs on VO_2 that is initially in the insulating state (324 K). These single-beam measurements show that the transmission at the *LC* resonance frequency (0.41 THz) increases with increasing incident field. Given the temperature-dependent data in Fig. 1c, this is consistent with an increase in the VO_2 conductivity in the SRR gaps. The higher-frequency dipole also redshifts as the in-gap VO_2 conductivity increases (peak at ~0.8 THz for the highest incident fields). At the highest field strength below damage threshold, the average conductivity increase is above 500 $(\Omega\,\mathrm{cm})^{-1}$, as estimated by the dipole frequency shift. Simulations assuming that the conductivity change occurs only within the horizontal gaps (Fig. 2e) agree well with experiment, although the measured resonance does not completely vanish. This is probably because the Gaussian-like THz beam profile leads to a larger effect in the centre of the beam than at the edge, where the THz field is weaker, and averaging of the measurement over the beam profile is not accounted for in the simulations.

The results in Fig. 2 clearly indicate a nonlinear response, but additional measurements are required to determine the dynamic response, which could suggest its probable microscopic origins. Figure 3 shows the time-dependent transmission of a weak-field THz probe pulse that was variably delayed with respect to the high-field THz pump pulse at an in-gap field strength of ~1 MV cm^{-1} (below the damage threshold). A plot of transmission as a function of frequency and pump–probe delay (Fig. 3a) shows the temporal evolution of the THz spectral response. From these data it is clear that the 0.41-THz *LC* resonance transmission increases, whereas the 1.0-THz dipole resonance exhibits a frequency shift. These changes occur on a picosecond timescale, as the VO_2 conductivity increases towards that of the metallic phase. Line-scans (Fig. 3b,c)more clearly reveal the dynamics. The time constants for initial change in transmission at 0.41 and 0.8 THz are 7 and 9 ps respectively, comparable to what has been observed in optical-pump, THz-probe experiments and consistent with a percolative phase transition[6,7,13]. The intrinsic conductivity response may be even faster, as the THz pump fields in the gaps are time-broadened (Fig. 2b). Nonetheless, it is clear that the THz pump induces a rapid change in the in-gap VO_2 conductivity. The change persists with little or no relaxation during the 100-ps range of our pump–probe time delay.

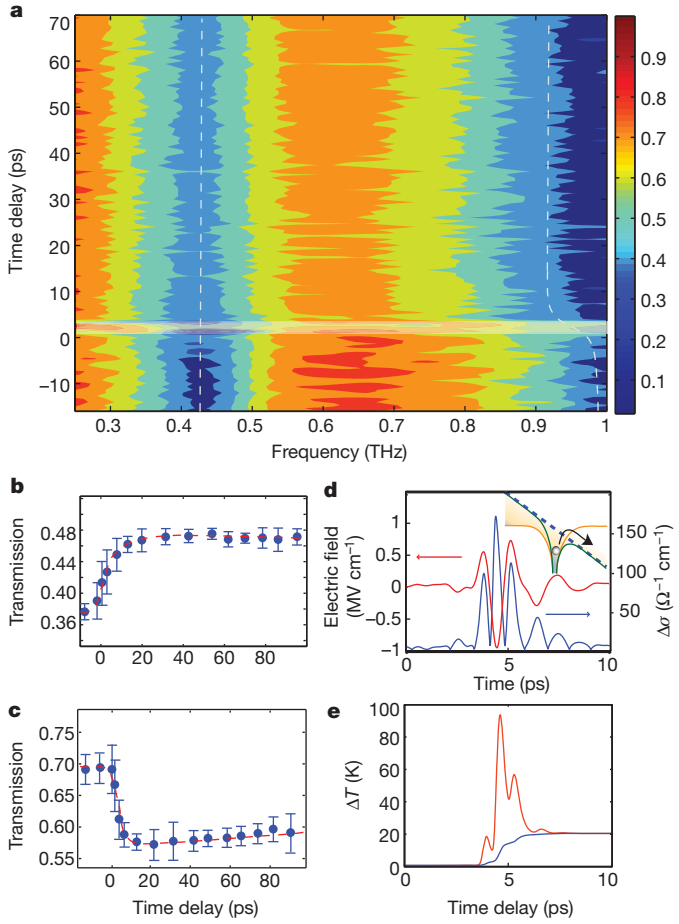


Figure 3 | THz pump–probe measurement and model calculation. **a**, Two-dimensional experimental data at 1 MV cm^{-1} in-gap field strength. The colour scale shows the transmission (unitless). **b** and **c**, Line-scans at (**b**) 0.41 THz, the metamaterial *LC* resonance, and (**c**) 0.8 THz, the metamaterial dipole resonance. Blue dots, data (324 K, 1 MV cm^{-1}); red curves, exponential fits. Error bars show standard error of the mean ($n = 8$). **d**, The in-gap electric field (red curve) results in a reduction of the confining potential (see inset), leading to the calculated Poole–Frenkel conductivity (blue curve). This releases carriers, leading to an absorbed power density that heats the electrons and subsequently the lattice, driving the VO_2 IMT. **e**, Electron (red) and lattice (blue) temperatures calculated using equation (2) (see text).

Our analysis suggests a two-step process (see Supplementary Information) for the THz-induced phase transition. First, the electric field reduces the Coulomb-induced activation barrier for carrier motion. This can be modelled by the Poole–Frenkel (PF) effect, described as[24,25]

$$\sigma = \sigma_0 \exp\left(\frac{\beta\sqrt{|E(t)|}}{rk_BT}\right) \quad (1)$$

where σ is the conductivity, σ_0 is the initial conductivity, $E(t)$ is the electric field, T is the temperature, and $\beta = (e^3/\pi\varepsilon)^{1/2}$ is the PF constant, where ε is the dielectric constant and r is a constant that depends on the position of the Fermi level. The PF effect contributes to the early dynamics of the IMT while the THz field is still acting on the sample. The electric field lowers the potential barrier to carrier hopping, increasing the carrier density[24–26]. A calculation of the transient PF conductivity is shown in Fig. 3d, along with a schematic illustration of the field-induced barrier reduction. The peak conductivity change ($\sim$150 (Ω cm)$^{-1}$) is consistent with the conductivity required to obtain the experimental transmission changes shown in Figs 2 and 3. In principle, if the carrier number density reaches the critical value ($\sim 10^{21}$ cm^{-3}) obtained by a modified Mott criterion for VO_2 (ref. 26), the PF effect alone would be sufficient to induce the IMT. However, as we now discuss, thermal effects rapidly follow the PF dynamics.

The PF-induced increase in carrier density serves as the initial condition for subsequent electric-field-induced carrier acceleration, leading to Joule heating through electron-lattice coupling[27]. This results in a temperature increase that drives the VO_2 into the persistent metallic phase. This can be modelled approximately with the well-known two-temperature model, describing the temporal evolution of the energy density in the electrons and lattice:

$$C_e\frac{dT_e}{dt} = -G(T_e - T_i) + \sigma(t)E^2(t)$$
$$C_i\frac{dT_i}{dt} = +G(T_e - T_i) \quad (2)$$

T_e (C_e) and T_i (C_i) are the temperature (specific heat) of the electrons and lattice, respectively, and G is the electron–phonon coupling coefficient, all of which have been determined experimentally for VO_2. The energy density from the incident electric field is approximated as $\sigma(t)E^2(t)$. Using the PF conductivity calculated in Fig. 3d and the experimental electric field (peak in-gap field 1 MV cm^{-1}), the above equation can be solved for T_e and T_i as a function of time, as shown in Fig. 3e. The initial electron heating dynamics, which approximately follow the THz intensity profile, are followed by equilibration of the lattice and electronic temperatures. The calculations indicate an increase of the VO_2 lattice temperature by $\sim$20 K on a several-picosecond timescale, consistent with the measured dynamics. The dynamics are clearly more complex than in our simple model, as the field enhancement decreases as the conductivity increases within the gaps. Nonetheless, THz-induced carrier release and acceleration followed by Joule heating on a picosecond timescale seems very likely to be the IMT

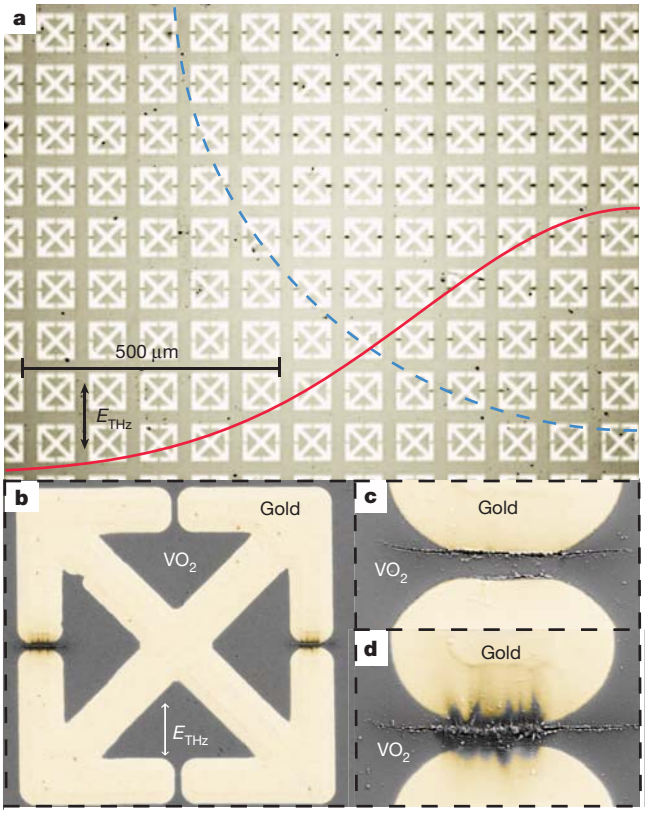


Figure 4 | THz-field-induced damage as revealed by optical and scanning electron micrographs. **a**, Millimetre-scale optical image of damage to VO_2 in the SRR gaps. The black dots are damaged VO_2. The dashed blue circle approximates the THz beam waist, and the red curve approximates the THz intensity profile. **b**, SEM image of a single SRR reveals that VO_2 is damaged by the vertically polarized THz field. **c** and **d**, Expanded view of damage at the edge of the THz beam (**c**) and near the beam centre (**d**). The gold SRRs were post-processed with false colour.

mechanism in our experiments. The large nonlinear transmission shown in Fig. 2 for the THz pump and in Fig. 3 for the THz probe is observed only within ~30 K of the transition temperature of 340 K, in the present experimental conditions. Note that a 1 MV cm^{-1} in-gap THz field corresponds to a THz fluence of ~2 mJ cm^{-2}, which is similar to the typical fluence threshold of the optically (800 nm) induced IMT[12,13,28]. The very different excitation wavelengths act initially on the system in very different ways, driving the short-time responses through different mechanisms. The extent of lattice heating that sustains the metallic phase at longer times may be comparable.

At our highest in-gap electric fields of ~4 MV cm^{-1} (~30 mJ cm^{-2} fluence), the THz electric field causes irreversible damage to the VO_2 metamaterials. Damage to the VO_2 thin film in the horizontal side gaps can be seen as black dots in Fig. 4a, with close-up images of the damage shown in Fig. 4b–d. The damage pattern depends strongly on the field strength, increasing towards the beam centre. The unique SRR geometry allows approximate visualization of the THz electric field, as the damage pattern follows the equipotential lines of the field (Supplementary Fig. 2).

In summary, we have demonstrated a THz-driven insulator–metal phase transition and shown that in VO_2 it is initiated by Poole–Frenkel electron liberation, followed by lattice equilibration on a picosecond timescale. Our work shows that metamaterial-enhanced high-field THz pulses can be used to study correlated-electron materials in a non-perturbative regime. The technique is extremely versatile, and can be used to study THz-induced phase transitions in other correlated materials and transition metal oxides (including high-T_c superconductors[29,30]), as well as THz-induced changes in electronic properties more generally. The metamaterial design can be optimized to balance the requirements in any particular measurement for maximum field enhancement, bandwidth and mode volume. Magnetic-field enhancement can also be studied, as SRRs provide temporal and spatial separation of the peak electric and magnetic fields.

METHODS SUMMARY

Metamaterial fabrication. Metamaterials on VO_2 films were made by stencil deposition techniques. The shadow masks used as stencils were made from 400-nm silicon nitride films with engraved metamaterial patterning. Once the masks were made, no photolithography was needed for the metamaterial fabrication onto the VO_2 surface. This prevented chemical contamination, thus ensuring high-quality samples.

High-field THz pulse generation and measurement. The output of a 1-kHz, 6.5-W, 100-fs Ti:sapphire amplifier was used to generate nearly single-cycle THz pulses by optical rectification in a lithium niobate crystal with the tilted pulse front technique. Our peak THz field strength was measured to be 300 kV cm^{-1}, with an estimated spot size of 1.5 mm. We used a standard electro-optic sampling setup to measure time-dependent fields. The THz field and a femtosecond optical pulse were overlapped spatially on a ZnTe electro-optic crystal, and the THz-induced optical birefringence was measured as the femtosecond pulse arrival time was varied, yielding the THz field temporal profile as in Fig. 2b (blue curve). Fourier transformation yielded the transmitted THz field in the frequency range 0.2–2.5 THz. Measurements with different THz pump–probe time delays yielded the two-dimensional data shown in Fig. 3a. All the experiments were performed in a high-vacuum cryostat with temperature control.

Electromagnetic simulations. The simulations in Fig. 2a–c, e and Supplementary Fig. 2d were performed using CST Microwave Studio 2011. All simulations used extremely fine mesh-cell settings, determined by adaptive meshing results (up to 8 million). All the parameters used in the CST simulations were those reported from experimental measurements; for example, the conductivity in the insulating state at 320 K is 10 $(\Omega\,\mathrm{cm})^{-1}$ and the relative permittivity of VO_2 in the insulating state is ~10. The simulations were performed using a time-domain transient solver.

Received 29 January; accepted 14 May 2012.
Published online 11 July 2012.

1. Morin, F. J. Oxides which show a metal-to-insulator transition at the Neel temperature. *Phys. Rev. Lett.* **3,** 34–36 (1959).
2. Limelette, P. *et al.* Universality and critical behavior at the Mott transition. *Science* **302,** 89–92 (2003).
3. Asamitsu, A., Tomioka, Y., Kuwahara, H. & Tokura, Y. Current switching of resistive states in magnetoresistive manganites. *Nature* **388,** 50–52 (1997).
4. Wang, J. *et al.* Epitaxial $BiFeO_3$ multiferroic thin film heterostructures. *Science* **299,** 1719–1722 (2003).
5. Cao, J. *et al.* Strain engineering and one-dimensional organization of metal-insulator domains in single-crystal vanadium dioxide beams. *Nature Nanotechnol.* **4,** 732–737 (2009).
6. Liu, M. K. *et al.* Photoinduced phase transitions by time resolved far-infrared spectroscopy in V_2O_3. *Phys. Rev. Lett.* **107,** 066403 (2011).
7. Cavalleri, A. *et al.* Femtosecond structural dynamics in VO_2 during an ultrafast solid-solid phase transition. *Phys. Rev. Lett.* **87,** 237401 (2001).
8. Berglund, C. N. & Guggenheim, H. J. Electronic properties of VO_2 near the semiconductor-metal transition. *Phys. Rev.* **185,** 1022–1033 (1969).
9. Stefanovich, G., Pergament, A. & Stefanovich, D. Electrical switching and Mott transition in VO_2. *J. Phys. Condens. Matter* **12,** 8837–8845 (2000).
10. Merbold, H., Bitzer, A. & Feurer, T. Second harmonic generation based on strong field enhancement in nanostructured THz materials. *Opt. Express* **19,** 7262–7273 (2011).
11. Chen, H.-T. *et al.* Active terahertz metamaterial devices. *Nature* **444,** 597–600 (2006).
12. Kübler, C. *et al.* Coherent structural dynamics and electronics correlations during an ultrafast insulator-to-metal phase transition in VO_2. *Phys. Rev. Lett.* **99,** 116401 (2007).
13. Hilton, D. J. *et al.* Enhanced photosusceptibility near T_c for the light-induced insulator-to-metal phase transition in vanadium dioxide. *Phys. Rev. Lett.* **99,** 226401 (2007).
14. Kim, H.-T. *et al.* Mechanism and observation of Mott transition in VO_2-based two- and three- terminal devices. *N. J. Phys.* **6,** 52 (2004).
15. Hoffmann, M. C., Hebling, J., Hwang, H. Y., Yeh, K.-L. & Nelson, K. A. THz-pump/THz-probe spectroscopy of semiconductors at high field strengths. *J. Opt. Soc. Am. B* **26,** A29–A34 (2009).
16. Yeh, K.-L., Hoffmann, M. C., Hebling, J. & Nelson, K. A. Generation of 10 μJ ultrashort THz pulses by optical rectification. *Appl. Phys. Lett.* **90,** 171121 (2007).
17. Hirori, H., Doi, A., Blanchard, F. & Tanaka, K. Single-cycle terahertz pulses with amplitudes exceeding 1 MV/cm generated by optical rectification in $LiNbO_3$. *Appl. Phys. Lett.* **98,** 091106 (2011).
18. Hoffmann, M. C., Hebling, J., Hwang, H. Y., Yeh K.-L. & Nelson K. A.. Impact ionization in InSb proved by terahertz pump-terahertz probe spectroscopy. *Phys. Rev. B* **79,** 161201 (2009).
19. West, K. G. *et al.* Growth and characterization of vanadium dioxide thin films prepared by reactive-based target ion beam deposition. *J. Vac. Sci. Technol. A* **26,** 133–139 (2008).
20. Werley, C. A. *et al.* Time-resolved imaging of near fields in THz antennas and direct quantitative measurement of field enhancements. *Opt. Express* **20,** 8551–8567 (2012).
21. Driscoll, T. *et al.* Memory metamaterials. *Science* **325,** 1518–1521 (2009).
22. Seo, M. A. *et al.* Terahertz field enhancement by a metallic nano slit operating beyond the skin-depth limit. *Nature Photon.* **3,** 152–156 (2009).
23. Shalaby, M. *et al.* Concurrent field enhancement and high transmission of THz radiation in nanoslit arrays. *Appl. Phys. Lett.* **99,** 041110 (2011).
24. Simmons, J. G. Poole-Frenkel effect and Schottky effect in metal-insulator-metal systems. *Phys. Rev.* **155,** 657–660 (1967).
25. Yeargan, J. R. & Taylor, H. L. The Poole-Frenkel effect with compensation present. *J. Appl. Phys.* **39,** 5600–5604 (1968).
26. Pergament, A., Boriskov, P. P., Velichko, A. A. & Kuldin, N. A. Switching effect and the metal-insulator transition in electric field. *J. Phys. Chem. Solids* **71,** 874–879 (2010).
27. Groeneveld, R. H. M., Sprik, R. & Lagendijk, A. Femtosecond spectroscopy of electron-electron and electron-phonon energy relaxation in Ag and Au. *Phys. Rev. B* **51,** 11433–11445 (1995).
28. Pashkin, A. *et al.* Ultrafast insulator-metal phase transition in VO_2 studied by multiterahertz spectroscopy. *Phys. Rev. B* **83,** 195120 (2011).
29. Basov, D. N. *et al.* Electrodynamics of correlated electron materials. *Rev. Mod. Phys.* **83,** 471–541 (2011).
30. Qazilbash, M. M. *et al.* Mott transition in VO_2 revealed by infrared spectroscopy and nano-imaging. *Science* **318,** 1750–1753 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge support from DOE-BES under grant DE-FG02-09ER46643 and from ONR grant N00014-09-1-1103.

Author Contributions R.D.A., K.A.N, M.L. and H.Y.H. came up with the experimental idea. H.Y.H. and M.L. performed the experiments. H.T., K.F., M.L., F.G.O. and X.Z. fabricated the metamaterial structures. A.J.S., M.L. and H.Y.H. performed full-wave electromagnetic simulation and analysed the data. K.G.W., S.K., J.L. and S.A.W. prepared the VO_2 thin films. A.C.S. and G.R.K. assisted with the simulation. M.L., H.Y.H., R.D.A. and K.A.N. wrote the manuscript. All authors contributed to the understanding of the underlying physics.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to. K.A.N. (kanelson@mit.edu) or R.D.A. (raveritt@buphy.bu.edu).

LETTER

doi:10.1038/nature11230

Dodecagonal tiling in mesoporous silica

Changhong Xiao[1]*, Nobuhisa Fujita[2]*, Keiichi Miyasaka[1,3], Yasuhiro Sakamoto[1,4] & Osamu Terasaki[1,3]

Recent advances in the fabrication of quasicrystals in soft matter systems have increased the length scales for quasicrystals[1] into the mesoscale range (20 to 500 ångströms). Thus far, dendritic liquid crystals[2], ABC-star polymers[3], colloids[4] and inorganic nanoparticles[5] have been reported to yield quasicrystals. These quasicrystals offer larger length scales than intermetallic quasicrystals (a few ångströms)[1,6], thus potentially leading to optical applications through the realization of a complete photonic bandgap induced via multiple scattering of light waves in virtually all directions[7–9]. However, the materials remain far from structurally ideal, in contrast to their intermetallic counterparts, and fine control over the structure through a self-organization process has yet to be attained. Here we use the well-established self-assembly of surfactant micelles to produce a new class of mesoporous silicas, which exhibit 12-fold (dodecagonal) symmetry in both electron diffraction and morphology. Each particle reveals, in the 12-fold cross-section, an analogue of dodecagonal quasicrystals in the centre surrounded by 12 fans of crystalline domains in the peripheral part. The quasicrystallinity has been verified by selected-area electron diffraction and quantitative phason strain analyses on transmission electron microscope images obtained from the central region. We argue that the structure forms through a non-equilibrium growth process, wherein the competition between different micellar configurations has a central role in tuning the structure. A simple theoretical model successfully reproduces the observed features and thus establishes a link between the formation process and the resulting structure.

Mesoporous silica, which is formed as a solid replica of a micellar aggregate, exhibits a large variety of structures. Two of the commonly observed crystal structures, with space groups $Pm\bar{3}n$ and $P4_2/mnm$, are of particular interest in connection with dodecagonal quasicrystals. Dodecagonal quasicrystals have been obtained in a dendritic liquid crystal system under synthesis conditions between those required to produce these two structures[2], so it seemed likely that quasicrystals could also be synthesized in mesoporous silica, which would be the first hard-matter quasicrystal in the mesoscale range. Here we explore an anionic surfactant system in which the two crystal structures (space groups $Pm\bar{3}n$ and $P4_2/mnm$) can be obtained merely by varying the alkalinity in the synthesis[10]. We find that an intermediate alkalinity leads to a crystal structure with the space group *Cmmm*. This new structure, however, is not the sole outcome of the synthesis, which also yields a structural arrangement that can be characterized as a dodecagonal quasicrystal[11].

All three crystal structures, with the space groups $Pm\bar{3}n$, $P4_2/mnm$ and *Cmmm*, are formed as tetrahedrally close-packed structures[12] of micelles (Fig. 1) encaged by silica, which remains in the final material and which represents the Voronoi tessellation for the micellar packing[13]. The $Pm\bar{3}n$ structure, corresponding to the A15-type structure of alloys[14], is described by two kinds of Voronoi polyhedron, namely

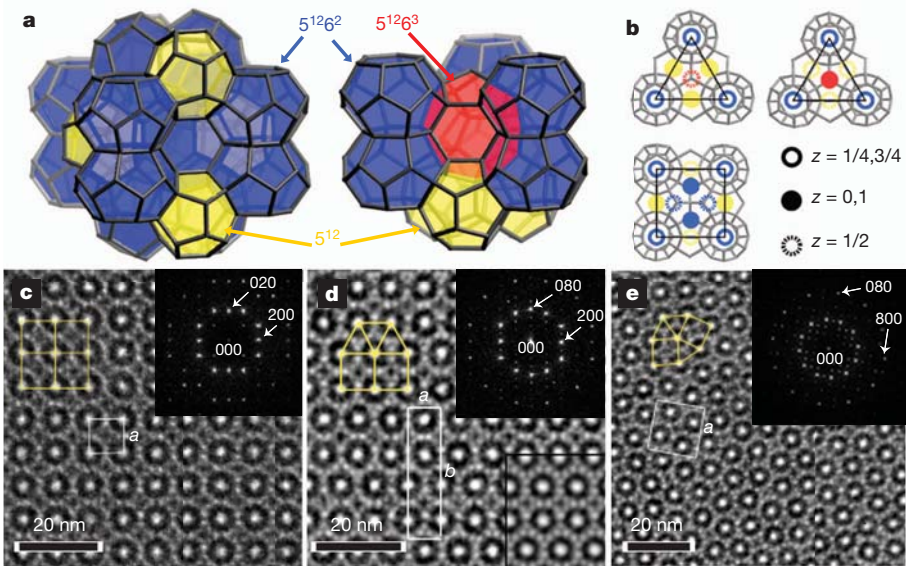


Figure 1 | The three basic crystal structures and their packing geometries. **a**, Three-dimensional packings for squares and triangles. The three kinds of Voronoi polyhedra, that is, $[5^{12}6^2]$, $[5^{12}6^3]$ and $[5^{12}]$, are shown blue, red and yellow, respectively. **b**, Projections of the polyhedral frames, demonstrating the two-dimensional/three-dimensional relationship of squares and triangles, where the centre of each polyhedron is indicated by a dot symbol (see key, where *z* stands for the fractional coordinate in the normal direction to the plane). **c–e,** TEM images and corresponding indexed diffraction patterns (top right insets) taken along the [001] axis for the $Pm\bar{3}n$ (**c**), *Cmmm* (**d**) and $P4_2/mnm$ (**e**) structures. The yellow and white lines in each TEM image show the corresponding tiling and the orthogonal unit cell. Bottom right inset in **d**, a simulated TEM image for the *Cmmm* structure.

[1]Department of Materials and Environmental Chemistry, Bezelii Center EXSELENT on Porous Materials, Stockholm University, S-10691 Stockholm, Sweden. [2]Institute of Multidisciplinary Research for Advanced Materials, Tohoku University, Sendai 980-8577, Japan. [3]Graduate School of EEWS (WCU), Korea Advanced Institute of Science and Technology, 335 Gwahangno, Yuseong-Gu, Daejeon 305-701, Republic of Korea (South Korea). [4]Nanoscience and Nanotechnology Research Center, Osaka Prefecture University, Sakai 599-8570, Japan.
*These authors contributed equally to this work.

$[5^{12}6^2]$ and $[5^{12}]$, where the $[5^{12}]$ polyhedra serve as the connection between columns of stacked $[5^{12}6^2]$ polyhedra along the [001] direction. (Here $[5^i6^j]$ denotes a polyhedron having i pentagonal and j hexagonal faces.) The $P4_2/mnm$ and $Cmmm$ structures, whose counterparts in alloys are the σ (ref. 15) and H (ref. 16) phases, respectively, require an additional kind of polyhedron, namely $[5^{12}6^3]$, to fill in the space among triangularly arranged $[5^{12}6^2]$ columns. The projected images of the $Pm\bar{3}n$, $P4_2/mnm$ and $Cmmm$ structures along the [001] direction can be represented as periodic tilings with squares and/or equilateral triangles. These are among the 11 Archimedean tilings[17] and are denoted by the vertex symbols 4^4, $3^2.4.3.4$ and $3^3.4^2$, respectively. These symbols represent the cyclic orders of squares (4) and/or equilateral triangles (3) surrounding each vertex, with the degeneracy of the identical polygons denoted by superscripts. Squares and equilateral triangles are also the elementary units of a dodecagonal quasiperiodic tiling (DQT)[18,19] that is used to model dodecagonal quasicrystals[11]. For a DQT, the ratio of the number of triangles to that of squares equals $4/\sqrt{3} \approx 2.31$ (ref. 20), and the tilings, $3^2.4.3.4$ and $3^3.4^2$, have twice as many triangles as squares.

We have performed a series of observations using transmission electron microscopy (TEM), which revealed that the $Cmmm$ structure (Fig. 1d) cannot be observed alone in a single particle. Instead, it always coexists with domains of the $Pm\bar{3}n$ and $P4_2/mnm$ structures and/or some modulated structures with squares and triangles, while retaining the periodicity along the common axis throughout the domains. More significantly, 12-fold-symmetric electron diffraction patterns are occasionally observed from the material, and a majority of the particles in the sample show the morphology of a dodecagonal prism (Fig. 2a, b), the size of which ranges from 1.5 μm to 4 μm.

To analyse the tiling characteristics within the whole cross-sectional area of the dodecagonal prism, we observed by TEM three thin slices (samples 1–3) prepared by ion beam thinning. The slices show a randomly tiled centre surrounded by 12 fans composed of more periodic tilings in the peripheral region (Fig. 2c, h). Each fan is a combination of rows of squares and equilateral triangles parallel to the edge of the dodecagon, forming $3^3.4^2$ and 4^4 vertices. Two adjacent fans that differ by 30° in orientation have a common narrow boundary area with $3^2.4.3.4$ vertices, defects and dislocated rows of squares and triangles (Fig. 2c). Frequently observed defects are shown in Fig. 2d–g. From a given defect a smaller fan-like arrangement may also grow, which can be incorporated into a major fan on either side. Hence, a fan may consist of several parallel domains bounded by defect rows (Fig. 2c and Supplementary Fig. 1). Although the arrangement can be seen as a two-dimensional analogue of a multiply twinned icosahedral particle consisting of face-centred cubic crystals[21], a clear contrast with the latter exists; there is no mirror symmetry across the boundary owing to a halfway shift of the rows of squares and triangles across the connecting units (typically, $3^2.4.3.4$ vertices).

At the centre, where the 12 fans meet, a disordered arrangement of tiles and defects is usually observed (Fig. 2h). The numbers of triangles and squares in the image (of area 350 nm × 350 nm) are 615 and 263, respectively; hence, the triangle/square ratio (~2.34) turns out to be reasonably close to that of a DQT (~2.31)[20]. The fast Fourier transform (FFT) of this image has almost perfect dodecagonal symmetry, except that the brightest peaks are elongated along a single direction (Fig. 2h).

Cross-sections of dodecagonal quasicrystals reported so far (including those mentioned at the beginning of this Letter) tend to show disorder in the arrangement of tiling units. The randomized tilings, however, maintain a 12-fold orientational order, and a set of sharp diffraction spots are often observed; such quasicrystals are called random tiling quasicrystals[22].

If the edge length is taken to be unity, every edge in a tiling of squares and equilateral triangles can be represented as one of six unit vectors, $\mathbf{e}_j = (\cos(\pi j/6), \sin(\pi j/6))$ with $j = 1$–6, the first four members of which constitute a linearly-independent basis set for the vertex coordinates. Hence, every vertex $\mathbf{r}$ is given as $\mathbf{r} = n_1\mathbf{e}_1 + n_2\mathbf{e}_2 + n_3\mathbf{e}_3 + n_4\mathbf{e}_4$, where

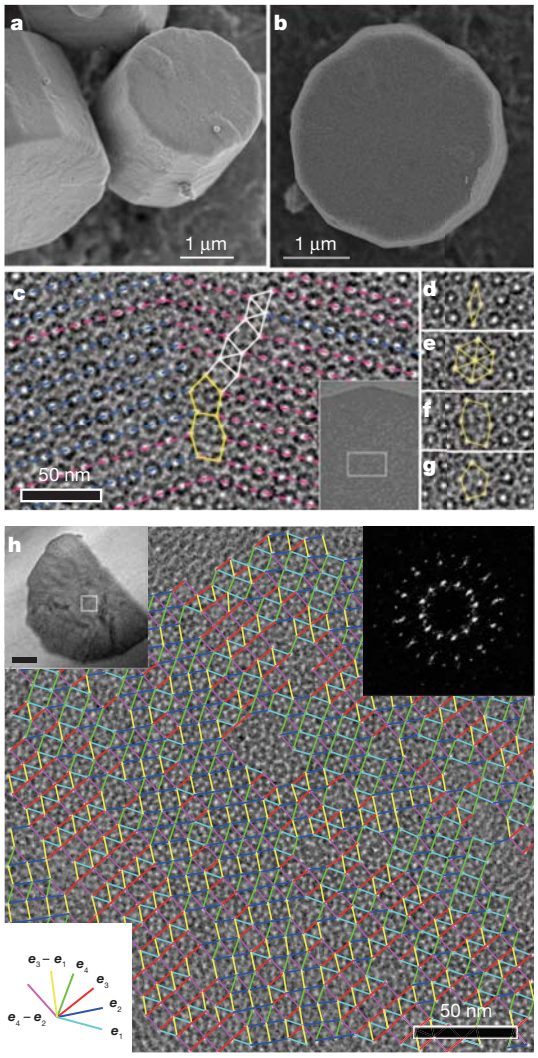


Figure 2 | Mesoporous particles with dodecagonal morphology and associated electron microscopy. **a**, **b**, SEM images taken from the sample; the particles are dodecagonal prisms. **c**, A TEM image taken from the peripheral part of sample 1. Bottom right inset, the box in the low magnification TEM image indicates from which part of the sample the tiling is observed. Irregular polygons (yellow lines) are observed in the boundary area between two fans. Linear rows of vertices within each fan are marked with broken lines: pink, $3^3.4^2$ vertices; light blue, 4^4 vertices. A defect row bounding two parallel domains is marked by white lines. **d**–**g**, Defects that are frequently observed between fans. **h**, Main panel, a TEM image taken from the central part of sample 2 (as indicated in the top left inset; scale bar, 1 μm), where the tiling edges are superposed on the image. The colours of the edges correspond to the six unit vectors in physical space as shown in the bottom left inset. Top right inset, FFT of the main-panel image.

n_j ($j = 1$–4) are integers that are uniquely determined. (This holds as long as the structure is free from any topological defect; see Supplementary Information for a related argument.) Although the set $\{\mathbf{e}_j\}_{j=1-4}$ is not a lattice basis in the plane, it is possible to define a lattice basis by doubling the dimensions via $\tilde{\mathbf{e}}_j = (\mathbf{e}_j, (-1)^j\mathbf{e}_j)$ ($j = 1$–4); that is, the set $\{\tilde{\mathbf{e}}_j\}_{j=1-4}$ generates the dodecagonal lattice (Λ_{dod}) in a four-dimensional hyperspace[23]. The first two components of the vector $\tilde{\mathbf{e}}_j$ belong to the physical space, while the last two belong to the orthogonal complement called the perpendicular space. It follows that every vertex $\mathbf{r}$ in a square–triangle tiling corresponds to a unique vertex $\boldsymbol{\rho} = n_1\tilde{\mathbf{e}}_1 + n_2\tilde{\mathbf{e}}_2 + n_3\tilde{\mathbf{e}}_3 + n_4\tilde{\mathbf{e}}_4$ in Λ_{dod}, and all the vertices when lifted into four dimensions constitute a so-called de Bruijn surface, which is a corrugated net embedded in Λ_{dod}. The tiling is an orthogonal projection of the net onto the physical space.

For a perfect DQT (for example, Stampfli tiling[19]), the net creeps along the physical space, with the perpendicular components $\mathbf{r}^{\perp} = \boldsymbol{\rho} - \mathbf{r}$ being bound within a finite region called a window[18,24] (also called an atomic surface or acceptance domain). When the net is on average inclined with respect to the physical space, the tiling is said to have a linear phason strain, which can be evaluated through a least-squares fitting of the coordinate data $\{\boldsymbol{\rho}\} = \{(\mathbf{r}, \mathbf{r}^{\perp})\}$ over the tiling to $\mathbf{r}^{\perp} \approx A\,\mathbf{r} + \mathbf{t}_0$, where the matrix A is called a phason strain tensor and $\mathbf{t}_0$ a phase shift. If A is non-zero, dodecagonal symmetry is no longer maintained in a strict sense[20]. The square root of the major eigenvalue of the matrix $A^{\mathrm{T}}A$, denoted λ, gives the magnitude of the linear phason strain, and the root mean square of the deviation, $|\mathbf{r}^{\perp} - (A\mathbf{r} + \mathbf{t}_0)|$, denoted d, measures the random phason disorder along the inclined direction. Thus, both λ and d provide quantitative measures, but of two different kinds, of the deviation of the structure from an ideal DQT.

The estimated value of d (1.640) over the tiling in Fig. 2h reflects the undulating nature of the de Bruijn surface or, equivalently, implies the significance of random phason disorder in the structure. The situation can be viewed somewhat differently: dividing the tiling into several domains can reduce the d value for each domain, so that the tiling is described as a random aggregate of periodic domains, including those of types $3^2.4.3.4$ ($d = 0.634$), $3^3.4^2$ ($d = 0.464$) and other associates called approximants. These domains have different configurations of squares and triangles, with λ values in the range between 0.268 (for $3^2.4.3.4$) and 1 (for $3^3.4^2$). The whole tiling is characterized by an even smaller linear phason strain ($\lambda = 0.211$), obtained by averaging the linear phasons of its constituents. (A detailed analysis is described in Supplementary Information.)

The randomness featured above is common to the tilings extracted from cross-sectional images of the particles. In most cases, the relevant dimensions exceed 200 nm × 200 nm, containing more than 400 units of squares and triangles. Observations made on non-sliced (that is, crushed) samples also demonstrate similar features. For example, sample 4 (Fig. 3c), obtained from the same sample batch as samples 1–3, gives $\lambda = 0.247$ and $d = 1.043$. On the other hand, smaller values of $\lambda = 0.124$ and $d = 0.916$ are found in sample 5 (Fig. 3d), which was obtained from another synthesis trial with slightly different synthesis conditions (Supplementary Information).

Besides the low values of λ and d, proximity to ideal quasicrystallinity can be demonstrated through diffraction experiments. An electron diffraction pattern taken from the centre of an ion sliced sample (sample 3) along the dodecagonal axis is shown in Fig. 3a. The pattern compares well to a simulated pattern calculated for a DQT. On the other hand, if the contributions from peripheral fans are included, some of the noticeable peaks fail to be indexed by a DQT, and many diffuse scattering features due to twinning are also observed. An extra peak, indicated in Fig. 3b, is associated with the Miller indices (400) for the *Cmmm* structure; the intensity of this peak, as well as that of other diffuse features, is reduced significantly when the aperture is narrowed. For samples 2, 4 and 5, 12-fold symmetry with sharp spots is demonstrated in Figs 2h, 3c and 3d, respectively, with an FFT of the TEM image or an electron diffraction pattern.

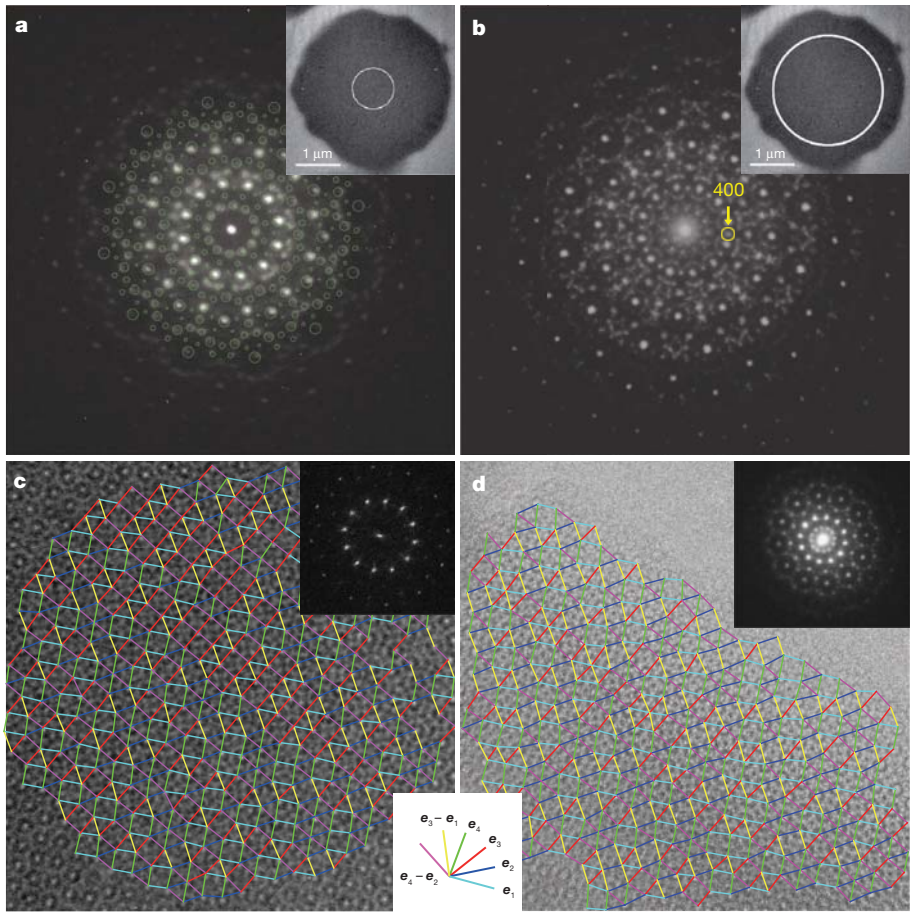


Figure 3 | Diffraction patterns and tilings indicating fine quasicrystallinity. **a**, **b**, Electron diffraction patterns taken from sample 3 using different aperture sizes, as indicated by the white circle in the respective insets. The diffraction intensities for a DQT calculated by assuming a point-wise scatterer at each vertex are shown by thin green curves on top in **a**, where the area of each circle corresponds to the intensity. In **b**, the peak marked with a yellow circle corresponds to the 400 reflection of the *Cmmm* structure. **c**, **d**, TEM images taken from two crushed quasicrystalline samples (samples 4 and 5, respectively). The colours of the edges superposed on the image correspond to the six unit vectors shown in the common inset. The inset at top right shows the FFT pattern for **c**, and the electron diffraction pattern for **d**. The numbers of triangles and squares in **c** and **d** are (382, 170) and (323, 142), respectively.

The peculiarity of the present material, with a quasicrystalline centre and a twin-like arrangement in the peripheral part of a dodecagonal-shaped particle, strongly suggests a non-equilibrium formation process. So far, the relative stabilities of different three-dimensional packings of soft spherical particles have been discussed in terms of an area-minimization principle for the interfaces[25–27]. Such theories analyse the equilibrium stability of the $Pm\bar{3}n$ structure, which corresponds to the simplest square–triangle tiling of interest, but do not go beyond this.

We now introduce a new approach to tackle the structural origins of the present material. Note that the three-dimensional structures can be abstracted as projected tilings in a plane, where a growing cluster at any instant is represented as a patch of a square–triangle tiling. The candidates for a new vertex are defined as all the positions that lie outside the patch and that are connected to any vertex on the outer boundary through one of the unit vectors $\mathbf{e}_j$, with $j = 1$–12. The occupation rate (p) for each candidate is given by $p = p_0[1 - \exp(\Delta E/T)]$ for $\Delta E < 0$ and $p = 0$ for $\Delta E \geq 0$, where ΔE is the increment of total energy on occupying the candidate vertex, and p_0 (the sampling rate) and T (temperature) serve as proper scales for the probability and energy, respectively. Growth proceeds by repeatedly making a random choice from the candidates with even probability, and by occupying a candidate vertex once it has been chosen a certain number of times, m. The value of m for every candidate with $p > 0$ is given as an integer that approximates $1/p$, whereas any candidate with $p = 0$ is rejected. This stochastic model is a modification of the Eden process[28,29], which was originally introduced to analyse the non-equilibrium growth of a colony of cancer cells or bacteria.

To mimic a real micellar system, ΔE is assumed to include the chemical potential, μ, and the interaction potential, V, gained when the candidate site is occupied. The following interaction potentials associated with different entities in the structure can be subsumed into V (Fig. 4a): (1) J_e, associated with every edge of the tiles; (2) J_r, with every right angle formed by two edges; (3) J_t or J_s, with every triangle or square, respectively; (4) J_{ss}, J_{st} or J_{tt}, with every contact between two tiles; and (5) J_{ttt}, with every cluster of three triangles ('s' and 't' in the subscripts stand for 'square' and 'triangle', respectively). In total, V is calculated as the sum of all the interaction terms for the entities newly generated by occupying the candidate site.

A survey of the above parameters sheds light on the primary factors that cause a systematic change in simulated structures. It turns out that unlimited aggregation of triangles can be avoided only if repulsive interactions are assumed between adjacent triangles; that is, J_{tt}, $J_{ttt} > 0$. Importantly, transformations from $3^2.4.3.4$ to a dodecagonal twin-like arrangement of $3^3.4^2$ and further to a quasicrystalline tiling are reproduced by reducing these repulsive interactions (Fig. 4), while keeping the other parameters constant. This is in accord with the experimental finding that these transformations occur through a small decrease in alkalinity, and thus each entity (edge, right angle, square, triangle) presumably maintains the same degree of stability.

The repulsions between triangles can possibly be attributed to the interfacial energy between the micelles. In a soft micellar aggregate, the total interfacial area can be reduced by expanding higher-coordinated cells against their lower-coordinated neighbours. Thus, if a cell having the highest coordination, $[5^{12}6^3]$, is surrounded only by lower-coordinated neighbours, it will expand to adjust its size so that the interfacial force is balanced by the micellar elasticity. If the adjacent neighbours also include highest-coordinated cells, which happens when triangles are in contact, the expansion is suppressed; therefore, a direct contact between triangles carries a cost in interfacial energy.

If a higher pH led to an increased electrical charge density of the surfactant heads and of the silica sources, the intra-layer Coulomb interactions would promote the stability as well as the flexibility of the interface layer[30]; the expansion of the highest-coordinated cells would thereby be facilitated, leading to enhanced repulsions between triangles. The $P4_2/mnm$ structure provides the highest stability at the highest pH. If the repulsions between triangles were suppressed by reducing the alkalinity, a specific formation that is seen at the tip of every fan-like domain of $3^3.4^2$ would arise, resulting in the growth of a multiply-twinned-like structure. Further reducing the alkalinity would maximize the competition between triangles and squares, leading to a random configuration of them. This is what we observe in the central part of our material. These adjustments require only a slight change of alkalinity; therefore, a fine-tuning of the alkalinity is at the heart of further improvement of the quasicrystallinity. Alternatively, $[5^{12}6^3]$ polyhedra would no longer be favoured by further reducing the alkalinity, as the $Pm\bar{3}n$ structure would then be superior.

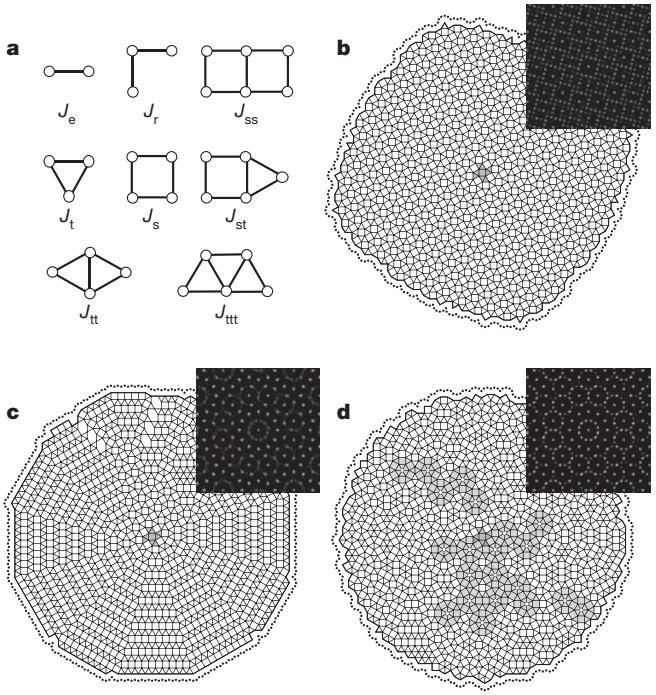


Figure 4 | Three tilings, each containing 1,500 vertices, simulated using different input parameters. **a**, A graphical presentation of the interaction potentials in the present model; see main text for details. **b**–**d**, Starting from the same screw-shaped patch in the centre of each panel, which is the dark grey shaded region composed of a square and eight triangles, each tiling is obtained with the following input parameters. Common parameters (see text for details) are $p_0 = 0.01$, $\mu/T = -0.0008$, $J_e/T = -0.0004$, $J_r/T = -0.0002$, $J_s/T = J_t/T = J_{ss}/T = 0$ and $J_{st}/T = -0.0008$. Parameters different for each panel are $(J_{tt}/T, J_{ttt}/T) = (0.0004, 0.00002)$ for **b**, (0.00027, 0) for **c**, and (0.00012, 0) for **d**. The boundary loops are somewhat emphasized by thicker curves, and the candidate sites are shown by dots. Panel **c** shows signs of the dodecagonal morphology with 12 fans. In **d**, the centres of dodecagonal wheel motifs, which are shaded light grey, are indicated by open dots. Connecting the dots with edges of length $2 + \sqrt{3}$ reveals larger squares and triangles, whose divisions demonstrate the inflation rules for the Stampfli tiling[19]. Insets at top right show the Fourier transform of each pattern. The numbers of triangles and squares in **b**, **c** and **d** are respectively (1,424, 709), (1,452, 689) and (1,507, 668).

Fine control of the micellar interactions through chemical conditions has enabled us to obtain quasicrystalline tilings in mesoporous silicas, where the non-equilibrium nature of the growth is essential for the formation. The present mechanism is expected to serve as a conceptual guide for synthesizing quasicrystals in a broader class of soft-matter systems in which the $Pm\bar{3}n$ and $P4_2/mnm$ structures are known to form.

METHODS SUMMARY

Synthesis. A method for synthesizing the $Pm\bar{3}n$ and $P4_2/mnm$ structures using an anionic surfactant, *N*-myristoyl-L-glutamic acid (C_{14}GluA), and a co-structure-directing agent (CSDA), *N*-trimethoxylsilylpropyl-*N*,*N*,*N*-trimethlammonium chloride (TMAPS), can be found in ref. 10. In the present work, the same method was used but with a series of intermediate NaOH concentrations. The present

material, characterized by the morphology of a dodecagonal prism, was obtained in one of the sample batches. More details on the synthesis are provided in Supplementary Information.

Characterization. High-resolution TEM was performed using a JEOL JEM-3010 microscope operating at 300 kV (coefficient of spherical aberration, Cs = 0.6 mm, point resolution 1.7 Å) and a JEOL JEM-2100 microscope operating at 200 kV (Cs = 1.4 mm, point resolution 2.5 Å). The morphological features were observed with a scanning electron microscope (SEM), JEOL JSM-7401F. Thinly sliced samples were prepared by an ion slicer, JEOL EM-09100 IS. The *Cmmm* structure was confirmed by comparing the relevant TEM images with a simulated TEM image from an idealized mesoporous *Cmmm* structure. The image simulation (Fig. 1d) was carried out using dedicated software, MesoPoreImage (http://web.mac.com/ohsuna/iWeb/E/Welcome.html). Additional details on the characterization are provided in Supplementary Information.

Tiling analysis. Tilings were extracted from TEM images, where tolerances of $\pm 10^{\circ}$ and $\pm 25\%$ are allowed for the edge orientation and the edge length, respectively. The obtained tilings were analysed using our original computer program. Details of the linear least-squares fitting of the coordinate data to evaluate the phason strain can be found in Supplementary Information.

Simulation. Extensions of the original Eden model were made so that growth over a dodecagonal support and the dependence of the growth rate on the local configuration could be handled. Simulations were performed with our original computer program. See the relevant descriptions in the text as well as in Supplementary Information.

Received 8 October 2011; accepted 8 May 2012.

1. Shechtman, D., Blech, I., Gratias, D. & Cahn, J. W. Metallic phase with long-range orientational order and no translational symmetry. *Phys. Rev. Lett.* **53,** 1951–1953 (1984).
2. Zeng, X. *et al.* Supramolecular dendritic liquid quasicrystals. *Nature* **428,** 157–160 (2004).
3. Hayashida, K., Dotera, T., Takano, A. & Matsushita, Y. Polymeric quasicrystal: mesoscopic quasicrystalline tiling in ABC star polymers. *Phys. Rev. Lett.* **98,** 195502 (2007).
4. Fischer, S. *et al.* Colloidal quasicrystals with 12-fold and 18-fold diffraction symmetry. *Proc. Natl Acad. Sci.* **108,** 1810–1814 (2011).
5. Talapin, D. V. *et al.* Quasicrystalline order in self-assembled binary nanoparticle superlattices. *Nature* **461,** 964–967 (2009).
6. Steurer, W. Twenty years of structure research on quasicrystals. Part I. Pentagonal, octagonal, decagonal and dodecagonal quasicrystals. *Z. Kristallogr.* **219,** 391–446 (2004).
7. Zoorob, M. E., Charlton, M. D. B., Parker, G. J., Baumberg, J. J. & Netti, M. C. Complete photonic bandgaps in 12-fold symmetric quasicrystals. *Nature* **404,** 740–743 (2000).
8. Man, W., Megens, M., Steinhardt, P. J. & Chaikin, P. M. Experimental measurement of the photonic properties of icosahedral quasicrystals. *Nature* **436,** 993–996 (2005).
9. Chan, Y. S., Chan, C. T. & Liu, Z. Y. Photonic band gaps in two dimensional photonic quasicrystals. *Phys. Rev. Lett.* **80,** 956–959 (1998).
10. Gao, C., Sakamoto, Y., Terasaki, O. & Che, S. Formation of diverse mesophases templated by a diprotic anionic surfactant. *Chem. Eur. J.* **14,** 11423–11428 (2008).
11. Ishimasa, T., Nissen, H. U. & Fukano, Y. New ordered state between crystalline and amorphous in Ni-Cr particles. *Phys. Rev. Lett.* **55,** 511–513 (1985).
12. Frank, F. C. & Kasper, J. S. Complex alloy structures regarded as sphere packings. II. Analysis and classification of representative structures. *Acta Crystallogr.* **12,** 483–499 (1959).
13. Sullivan, J. M. in *Foams and Emulsions* (eds Sadoc, J. F. & Rivier, N.) 379–402 (Kluwer Academic, 1998).
14. Borén, B. Röntgenuntersuchung der Legierungen von Silicium mit Chrom, Mangan, Kobalt und Nickel. *Ark. Kemi. Miner. Geol.* **11,** 1–28 (1933).
15. Bergman, G. & Shoemaker, D. P. The determination of the crystal structure of the sigma phase in the iron-chromium and iron-molybdenum systems. *Acta Crystallogr.* **7,** 857–865 (1954).
16. Ye, H. Q., Li, D. X. & Kuo, K. H. Structure of the H phase determined by high-resolution electron microscopy. *Acta Crystallogr. B* **40,** 461–465 (1984).
17. Grunbaum, B. & Shephard, G. C. *Tilings and Patterns* (Freeman, 1986).
18. Baake, M., Klitzing, R. & Schlottmann, M. Fractally shaped acceptance domains of quasiperiodic square-triangle tilings with dodecagonal symmetry. *Physica A* **191,** 554–558 (1992).
19. Stampfli, P. A dodecagonal quasiperiodic lattice in two dimensions. *Helv. Phys. Acta* **59,** 1260–1263 (1986).
20. Leung, P. W., Henley, C. L. & Chester, G. V. Dodecagonal order in a two-dimensional Lennard-Jones system. *Phys. Rev. B* **39,** 446–458 (1989).
21. Miyasaka, K., Han, L., Che, S. & Terasaki, O. A lesson from the unusual morphology of silica mesoporous crystals: growth and close packing of spherical micelles with multiple twinning. *Angew. Chem.* **118,** 6666–6669 (2006).
22. Oxborrow, M. & Henley, C. L. Random square-triangle tilings: a model for twelvefold-symmetric quasicrystals. *Phys. Rev. B* **48,** 6966–6998 (1993).
23. Yamamoto, A. Crystallography of quasiperiodic crystals. *Acta Crystallogr. A* **52,** 509–560 (1996).
24. Cockayne, E. Nonconnected atomic surfaces for quasicrystalline sphere packings. *Phys. Rev. B* **49,** 5896–5910 (1994).
25. Ziherl, P. & Kamien, R. D. Maximizing entropy by minimizing area: towards a new principle of self-organization. *J. Phys. Chem. B* **105,** 10147–10158 (2001).
26. Weaire, D. & Phelan, R. A counter-example to Kelvin's conjecture on minimal surfaces. *Phil. Mag. Lett.* **69,** 107–110 (1994).
27. Kusner, R. & Sullivan, J. M. in *The Kelvin Problem: Foam Structures of Minimal Surface Area* (ed. Weaire, D.) 71–80 (Taylor and Francis, 1996).
28. Eden, M. in *Symposium on Information Theory in Biology* (ed. Yockey, P. H.) 359–370 (Pergamon, Symposium Publications Division, 1958).
29. Meakin, P. Noise-reduced and anisotropy-enhanced Eden and screened-growth models. *Phys. Rev. A* **38,** 418–426 (1988).
30. Durian, D. J. & Raghavan, S. R. Making a frothy shampoo or beer. *Phys. Today* **63,** 62–63 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank K. Niizeki, T. Dotera, A. E. Garcia-Bennett, S. Che and C. Gao for discussions, O. M. Yaghi and M. O'Keeffe for critical reading of the manuscript, and J. Shen for encouragement and support. This work was supported by the Swedish Research Council (VR), the Japan Science and Technology Agency (JST) and Berzelii EXSELENT. SEM and TEM studies were performed at the Electron Microscopy Center (EMC) at Stockholm University, which is supported by the Knut and Alice Wallenberg Foundation. Support from the WCU programme, Korea (R-31-2008-000-10055-0; K.M. and O.T.), Grants-in-Aid for Young Scientists (B) of JSPS (no. 23710132; Y.S.), and Special Coordination Funds for Promoting Science and Technology of MEXT, Japan (Y.S.) is also acknowledged.

Author Contributions C.X. synthesized the materials and carried out electron microscopy observations. C.X. and N.F. analysed the tilings obtained experimentally. N.F. developed the theoretical part, including the modelling of the energetics and the growth process. K.M. and Y.S. contributed early TEM observations and data analysis. C.X. and N.F. wrote the manuscript with inputs from all co-authors. O.T. initiated and led the project. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to O.T. (terasaki@mmk.su.se).

LETTER

doi:10.1038/nature11294

Solid–liquid iron partitioning in Earth's deep mantle

Denis Andrault[1], Sylvain Petitgirard[2], Giacomo Lo Nigro[1], Jean-Luc Devidal[1], Giulia Veronesi[2], Gaston Garbarino[2] & Mohamed Mezouar[2]

Melting processes in the deep mantle have important implications for the origin of the deep-derived plumes believed to feed hotspot volcanoes such as those in Hawaii[1]. They also provide insight into how the mantle has evolved, geochemically and dynamically, since the formation of Earth[2]. Melt production in the shallow mantle is quite well understood, but deeper melting near the core–mantle boundary remains controversial. Modelling the dynamic behaviour of deep, partially molten mantle requires knowledge of the density contrast between solid and melt fractions. Although both positive and negative melt buoyancies can produce major chemical segregation between different geochemical reservoirs, each type of buoyancy yields drastically different geodynamical models. Ascent or descent of liquids in a partially molten deep mantle should contribute to surface volcanism or production of a deep magma ocean, respectively. We investigated phase relations in a partially molten chondritic-type material under deep-mantle conditions. Here we show that the iron partition coefficient between aluminium-bearing (Mg,Fe)SiO_3 perovskite and liquid is between 0.45 and 0.6, so iron is not as incompatible with deep-mantle minerals as has been reported previously[3]. Calculated solid and melt density contrasts suggest that melt generated at the core–mantle boundary should be buoyant, and hence should segregate upwards. In the framework of the magma oceans induced by large meteoritic impacts on early Earth, our results imply that the magma crystallization should push the liquids towards the surface and form a deep solid residue depleted in incompatible elements.

Zones in the lowermost mantle where seismic waves have ultra-low velocities have been interpreted as indicative of partial melting[1]. Between 5% and 30% partial melting could reduce the velocities of P and S seismic waves by between 10% and 30% for specific geometries of liquid domains in the solid–liquid mixture, in agreement with seismological observations[4,5]. Such values represent a high degree of partial melting in comparison with the approximately 10–15% observed in the shallow mantle beneath mid-ocean ridges. Furthermore, this value for ridges is the result of an integration over a depth interval of 30–40 km of locally very low degrees of partial melting: less than 1% (ref. 6). If partial melting were that high in the lowermost mantle, the consequences for the formation of distinct geochemical reservoirs would be major[7]. Deep-mantle seismic features could also be attributable to heterogeneities in the purely solid phases, produced by core–mantle chemical interactions or the presence of foreign materials introduced by large-scale mantle convection[8,9]. Although recent estimations of solidus melting temperature converge on 4,150 K at the core–mantle boundary (CMB) pressure of 135 GPa (refs 10, 11), our knowledge of the temperature in this region remains too vague to assert melting in the lowermost mantle.

The major parameters controlling the buoyancy of deep-mantle melts are: the changes in atomic packing between solid and liquid phases that can contribute to a volume increase of 3–4% (refs 12, 13); the iron partitioning coefficient between solid and liquid mantle fractions; and the MgO/SiO_2 ratio in the liquid phase[14]. We carried out experiments to determine these parameters more accurately, using synthetic (Ca,Mg,Al,Si,Fe) glass with a composition that modelled the primitive mantle after core segregation (similar to pyrolite, and the same sample as used previously[10]). We did not include minor and trace elements, because their effect on iron partitioning can be neglected to a first approximation. We did the experiments in a diamond-anvil cell (DAC) at pressures from 40 GPa to 120 GPa, using very thin (5–10 μm) samples that were melted throughout their thickness using infrared lasers (Fig. 1). Melting criteria are based on the use of *in situ* X-ray diffraction, and details are reported elsewhere[10]. We analysed the recovered samples using simultaneous X-ray diffraction (XRD) and X-ray fluorescence (XRF) (Fig. 2), generating XRD and XRF maps with resolutions up to around 500 nm (Table 1). We successfully extracted some samples from their gaskets, and analysed them chemically using the electron microprobe (see Methods and Supplementary Information).

The XRD maps (Fig. 3b, d) reveal circular zoning around the centre of the laser hotspot (CLHS). Multiphase Rietveld refinements of each XRD pattern provide maps of phase contents for each pixel in the two-dimensional maps. XRF spectra recorded at the same sample positions

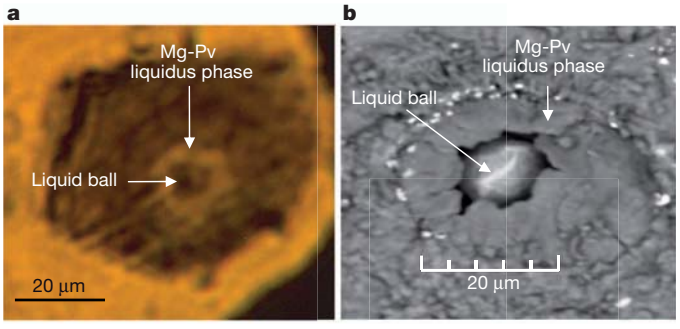


Figure 1 | **a**, Optical and **b**, scanning-electron micrographs of samples recovered after partial melting at high pressure. Sample (**a**) was heated for a few seconds at 3,650 K and 78.5 GPa. Sample (**b**) was heated for about one minute at 3,200 K and 55 GPa.

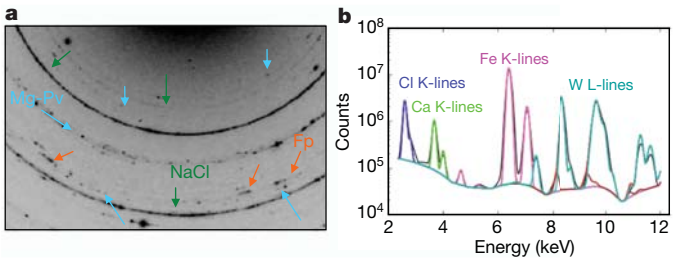


Figure 2 | Examples of raw experimental data. After samples were melted and pressure released, we took XRD measurements at several sample positions. Typical spectrum (**a**) shows a mixture of Mg-Pv (blue), Fp (orange) and the NaCl (green) pressure medium. The calcium silicate perovskite is not visible in this sample region. We also recorded XRF measurements at the same sample positions (**b**). Visible peaks arise from Cl in NaCl (blue), Ca (light green) and Fe (pink) in the sample and W (dark green) in the gasket. Refinement of XRF spectra provides quantitative maps of Fe content and qualitative maps for the lighter elements (See Supplementary Information).

[1]Laboratoire Magmas et Volcans, Université Blaise Pascal, CNRS, IRD, 63038 Clermont-Ferrand, France. [2]European Synchrotron Radiation Facility, 38043 Grenoble, France.

Table 1 | Experimental conditions and results summary for the nine experiments performed in this study.

Pressure (GPa) at 300 K	Liquidus temperature (K)	Pressure (GPa) at high temperature	Highest map resolution (μm^2)	Iron partition coefficient (perovskite/liquid)	Standard deviation
35	2,900	41.5	<0.5 × 0.5>	0.60	0.08
41	3,050	48.0	[2 × 2]	0.56	0.09
50	3,250	57.5	[2.5 × 2.5]	0.55	0.10
60.5	3,450	68.5	[2 × 2]	0.50	0.09
70	3,650	78.5	<0.5 × 0.5>	0.50	0.08
75	3,750	83.5	[2 × 2]	0.56	0.06
95	4,150	105.0	<0.5 ×0.5 >	0.54	0.08
103	4,250	113.0	[1.5 × 1.5]	0.57	0.07
110	4,400	120.0	<0.5 × 0.5>	0.47	0.09

More information on the sample conditions is reported in the Supplementary Information. Square brackets, XRF maps performed at the ID27 beamline; angle brackets, ID21 beamline.

show chemical zoning with a circular shape around the CLHS, similar to the XRD features (Figs 3a, c and Supplementary Fig. 1). The amplitude of chemical heterogeneities is illustrated by profile analyses of the XRF intensities at the iron K-lines (Supplementary Figs 2–5). Comparison of XRD and XRF maps shows that zones with a greater content of magnesium-bearing perovskite (Mg-Pv) do not overlap with zones with higher iron content. This feature indicates that iron is relatively incompatible with Mg-Pv. Our samples adopt a geometrical configuration similar to those obtained in other high-pressure devices, such as the large-volume press[15–18]. The liquid phase is found at higher temperatures (that is, at the CLHS), whereas the liquidus-phase Mg-Pv is found at the interface with the quenched liquid (Fig. 1). Our sample shape is also perfectly compatible with that observed in a previous laser-heated DAC study done under comparable pressure–temperature conditions[3].

The iron partition coefficient between melt and solid at the liquidus temperature ($D_{Fe} = X_{Fe}^{sol} / X_{Fe}^{liq}$, where X_{Fe}^{sol} and X_{Fe}^{liq} are the molar iron contents in the solid and liquid phases, respectively) is inferred from variations in iron content at different sample positions. To calculate D_{Fe}, we combined maps of mineralogical and chemical contents, extracted from XRD and XRF respectively. D_{Fe} ranges from 0.60(5) to 0.47(5) for nominal pressures of 35 GPa and 110 GPa, respectively (Table 1, Fig. 4). This confirms that iron is relatively incompatible with the solid mantle. Our D_{Fe} values at the lowest pressures are at the higher end of a number of experimental results obtained using multi-anvil press apparatus (see Supplementary Information). They are also in very good agreement with two previous studies[19,20].

Concerning the deep mantle, our data set plots at significantly higher D_{Fe} values than have previously been reported[3]. A major difference comes from the composition of the starting materials: olivine in the previous study[3] and an aluminium-bearing silicate glass in the present one. At a pressure of around 65 GPa, for example, the discrepancy in D_{Fe} between the two studies accounts for a change from around 0.2 to around 0.6, for aluminium-free and aluminium-bearing samples respectively. The presence of aluminium could explain the systematic difference in the results, because aluminium is known to enhance the incorporation of iron into perovskite. At 25 GPa, K_{Fe}, the iron partition coefficient between Mg-Pv and ferropericlase (Fp), was reported to change from 0.36 to 1 when aluminium is present in the starting material[21]. Similarly, K_{Fe} between Fp and the post Mg-Pv phase was reported to vary from 0.15 (ref. 22) or 0.4 (ref. 23) to about 4 (refs 24, 25) for aluminium-free and aluminium-bearing systems, respectively. On this basis, the iron partition coefficients reported in the olivine system[3] are not very relevant to melting of Earth's deep mantle.

Our results indicate a liquid 1.6–2.1 times richer in iron than the Mg-Pv phase under pressure conditions similar to the lower mantle. On this basis, we can model the liquid–solid phase relations in partially melted deep mantle. A liquid in chemical equilibrium with a pyrolitic or chondritic lower mantle with FeO of around 11 mol% should adopt an

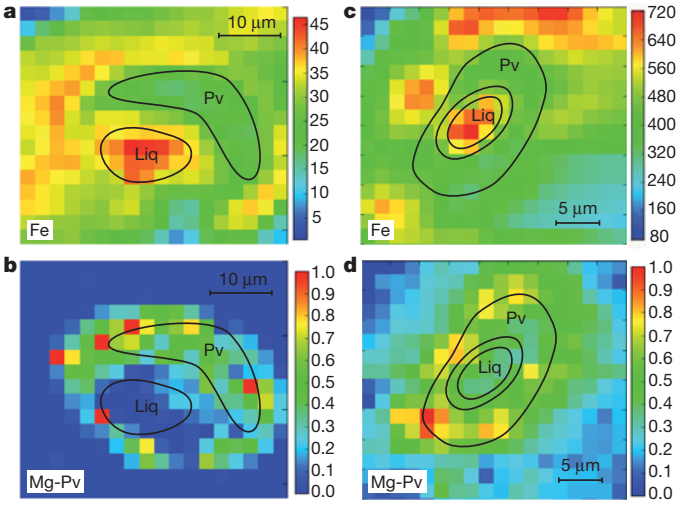


Figure 3 | Spatial distributions of Fe XRF intensity (**a**, **c**) and Al-bearing Mg-Pv contents (**b**, **d**) for samples synthesized at 57.5 GPa (**a**, **b**) and 113 GPa (**c**, **d**). Colours represent thousands of counts for Fe fluorescence and are normalized to 1 for Mg-Pv content. Samples present clear chemical and mineralogical zoning. Zones with maximum Fe concentration do not overlap with zones of maximum Mg-Pv concentration. The Fe fluorescence intensity is highest at the centre of the laser hotspot, where the liquid phase is concentrated. We derive the degree of Fe incompatibility in the solid from the contrast in Fe content between the different sample zones. Black lines superimposed on the figure delineate the regions in which XRF intensities have been averaged for calculating the relative Fe contents in the liquid and liquidus phases (see Methods).

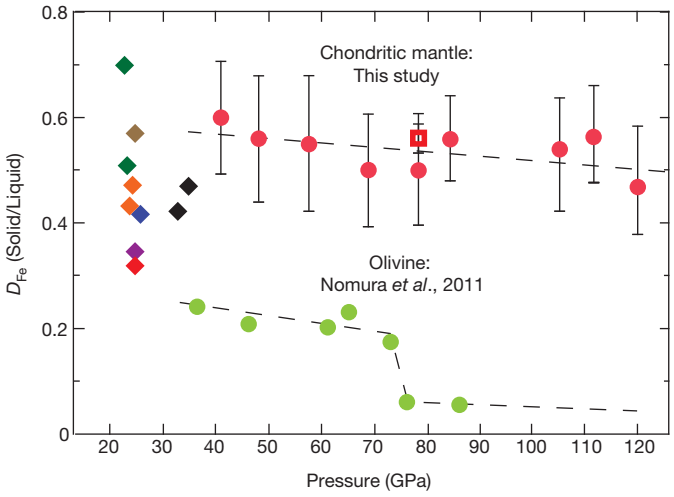


Figure 4 | Pressure evolution of the Fe partition coefficient (D_{Fe}) between silicate melt and the Mg-Pv liquidus phase, determined using XRF analyses (red circles) and EPMA electron-probe micro-analyses (red square) of this study. Our error bars represent the standard deviation (see Table 1). We also report recent results obtained after melting of olivine (light-green circles, ref. 3). The coloured diamonds correspond to previous data sets obtained using the large volume press in previous studies: ref. 19 (brown), ref. 28 (purple), ref. 15 (black), ref. 20 (dark green), ref. 29 (red), ref. 17 (orange), ref. 30 (blue).

FeO content of 15–20 mol%. On the basis of thermodynamic calculations presented in a recent study[14], we calculate that a liquid with 20 mol% FeO is buoyant in the lowermost solid mantle if the liquid SiO_2 content is higher than 20 mol%. Previous studies reported liquid SiO_2 contents of about 40 mol% at 33 GPa (ref. 15) and about 30 mol% at 135 GPa (ref. 3). Therefore, 20 mol% SiO_2 is too low to be achieved at lower-mantle conditions and our results for D_{Fe} support buoyant silicate liquids at all mantle depths. An increased temperature might even accentuate the upward movements, if the liquids are generated in contact with the relatively hotter CMB[26].

We reach a similar conclusion when other geological circumstances are considered. In early Earth, complete melting of a large fraction of the deep mantle would yield a liquid composition close to that of the bulk-silicate mantle: that is, with an FeO-content of about 11 mol%. Progressive crystallization of this liquid should produce an Mg-Pv phase with an FeO content of about 6 mol%, based on the D_{Fe} value of 0.55. When compared with the situation described in the previous paragraph, this solid–liquid equilibrium results in a decrease of FeO content in the liquid (from 20 mol% to 11 mol%) that is significantly larger than the decrease in the solid (from 11 mol% to 6 mol%). It would certainly contribute to a larger liquid buoyancy. Accordingly, it is expected that the crystallizing Mg-Pv grains should sink in the magma ocean. The arrival in the deepest part of the magma ocean of such material depleted in incompatible elements should make its crystallization easier, because the melting temperature is increased. This effect, together with the shape of melting curves in the deep mantle[10], is consistent with crystallization of the magma ocean starting from the bottom and progressing to the top, during the release of heat at the mantle surface. This can produce a vertical layering of Earth's mantle, in agreement with geodynamic modelling of the magma ocean[2].

A large amount of liquid being trapped in a basal magma ocean for long geological times in the very deep mantle[27] seems to be incompatible with our experimental results. We think that such liquids should eventually have risen. Also, the origin of such deep liquids can hardly be related to the early occurrence of magma oceans at Earth's surface, because silicate liquids are not expected to sink towards the CMB. Consequently, such a dense basal magma ocean, if it existed, should be preferentially related to the existence of a very hot core, inducing partial melting in the lowermost mantle. This situation could have persisted for long geological times, independent of the episodic mantle melting induced by meteoritic impacts at Earth's surface.

METHODS SUMMARY

Very thin (5–10 μm) samples were loaded into the DAC between two NaCl pellets of 5–10 μm thickness each. The use of relatively thick NaCl pellets minimizes the axial temperature gradient and allows the sample to melt throughout its thickness. After compression to the target pressure, samples were heated by two infrared lasers with a spot diameter of more than 20 μm. Temperature was measured by analysing the pyrometric signal emitted over a sample area of 3 μm × 3 μm. Sample melting was achieved for durations ranging from a couple of seconds to around 1 min before rapid quenching.

For the diffraction measurements, the X-ray beam of the ID27 beamline at the European Synchrotron Radiation Facility (ESRF) was tuned to 33 keV and focused by two Kirkpatrick–Baez mirrors to less than 2 μm × 2 μm on the sample. We used the MAR-160 charge-coupled-device detector with a typical acquisition time of 20–30 s. Multiphase Rietveld refinements provided phase contents of Mg-Pv and Ca-bearing perovskite and Fp. The diffraction signal decreases away from the CLHS, where the glassy starting material was not heated enough for it to crystallize.

X-ray fluorescence analysis was done at the ESRF's ID27 and ID21 beamlines. All reported measurements were taken after release of pressure and temperature. We used different geometries with the detector located at about 60° or 90° from the incident beam. For some experiments, we used a polycapillary X-ray half-lens to enhance the XRF signal–noise ratio. At ID21, the monochromatic beam was tuned to 7.2 keV and focused to 0.2 μm × 0.7 μm using a Fresnel Zone Plate lens. Typical acquisition times were 100–150 s per spectrum at ID27 and 1 s per spectrum at ID21. We detected K-lines of Ca and Fe from the sample, the K-line of Cl from the NaCl pellets and L-lines of W or Re from the gasket material. We could not measure Mg, Si and Ca fluorescence with sufficient accuracy to draw major consequences for their repartition within the melted regions.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 15 July 2011; accepted 31 May 2012.

1. Lay, T., Garnero, E. J. & Williams, Q. Partial melting in a thermo-chemical boundary layer at the base of the mantle. *Phys. Earth Planet. Inter.* **146,** 441–467 (2004).
2. Solomatov, V. S. in *Origin of the Earth and Moon* (eds Canup, R. M. & Righter, K.) 323–338 (Univ. Arizona Press, 2000).
3. Nomura, R. *et al.* Spin crossover and iron-rich silicate melt in the Earth's deep mantle. *Nature* **473,** 199–202 (2011).
4. Hernlund, J. W. & Jellinek, A. M. Dynamics and structure of a stirred partially molten ultralow-velocity zone. *Earth Planet. Sci. Lett.* **296,** 1–8 (2010).
5. Rost, S., Garnero, E. J., Williams, Q. & Manga, M. Seismological constraints on a possible plume root at the core–mantle boundary. *Nature* **435,** 666–669 (2005).
6. Laporte, D., Toplis, M. J., Seyler, M. & Devidal, J. L. A new experimental technique for extracting liquids from peridotite at very low degrees of melting: application to partial melting of depleted peridotite. *Contrib. Mineral. Petrol.* **146,** 463–484 (2004).
7. McNamara, A. K., Garnero, E. J. & Rost, S. Tracking deep mantle reservoirs with ultra-low velocity zones. *Earth Planet. Sci. Lett.* **299,** 1–9 (2010).
8. Buffett, B. A., Garnero, E. J. & Jeanloz, R. Sediments at the top of Earth's core. *Science* **290,** 1338–1342 (2000).
9. Sakai, T. *et al.* Interaction between iron and post-perovskite at core-mantle boundary and core signature in plume source region. *Geophys. Res. Lett.* **33,** L15317 (2006).
10. Andrault, D. *et al.* Melting curve of the deep mantle applied to properties of early magma ocean and actual core-mantle boundary. *Earth Planet. Sci. Lett.* **304,** 251–259 (2011).
11. Fiquet, G. *et al.* Melting of peridotite to 140 gigapascals. *Science* **329,** 1516–1518 (2010).
12. Mosenfelder, J. L., Asimow, P. D., Frost, D. J., Rubie, D. C. & Ahrens, T. J. The $MgSiO_3$ system at high pressure: thermodynamic properties of perovskite, postperovskite, and melt from global inversion of shock and static compression data. *J. Geophys. Res.* **114,** B01203 (2009).
13. Stixrude, L., de Koker, N., Sun, N., Mookherjee, M. & Karki, B. B. Thermodynamics of silicate liquids in the deep Earth. *Earth Planet. Sci. Lett.* **278,** 226–232 (2009).
14. Funamori, N. & Sato, T. Density contrast between silicate melts and crystals in the deep mantle: an integrated view based on static-compression data. *Earth Planet. Sci. Lett.* **295,** 435–440 (2010).
15. Ito, E., Kubo, A., Katsura, T. & Walter, M. J. Melting experiments of mantle materials under lower mantle conditions with implications for magma ocean differentiation. *Phys. Earth Planet. Inter.* **143–144,** 397–406 (2004).
16. Litasov, K. & Ohtani, E. Phase relations and melt compositions in CMAS–pyrolite–H2O system up to 25 GPa. *Phys. Earth Planet. Inter.* **134,** 105–127 (2002).
17. Trønnes, R. G. & Frost, D. J. Peridotite melting and mineral–melt partitioning of major and minor elements at 22–24.5 GPa. *Earth Planet. Sci. Lett.* **197,** 117–131 (2002).
18. Zhang, J. & Herzberg, C. Melting experiments on anhydrous peridotite KLB-1 from 5.0 to 22.5 GPa. *J. Geophys. Res.* **99,** 17729–17742 (1994).
19. Ohtani, E., Moriwaki, K., Kato, T. & Onuma, K. Melting and crystal–liquid partitioning in the system Mg_2SiO_4–Fe_2SiO_4 to 25 GPa. *Phys. Earth Planet. Inter.* **107,** 75–82 (1998).
20. Walter, M. J., Nakamura, E., Trønnes, R. G. & Frost, D. J. Experimental constraints on crystallization differentiation in a deep magma ocean. *Geochim. Cosmochim. Acta* **68,** 4267–4284 (2004).
21. Wood, B. J. & Rubie, D. C. The effect of alumina on phase transformations at the 660-kilometer discontinuity from Fe–Mg partitioning experiments. *Science* **273,** 1522–1524 (1996).
22. Auzende, A.-L. *et al.* Element partitioning between magnesium silicate perovskite and ferropericlase: new insights into bulk lower-mantle geochemistry. *Earth Planet. Sci. Lett.* **269,** 164–174 (2008).
23. Kobayashi, Y. *et al.* Fe–Mg partitioning between (Mg, Fe)SiO_3 post-perovskite, perovskite and magnesiowüstite in the Earth's lower mantle. *Geophys. Res. Lett.* **32,** L19301 (2005).
24. Andrault, D. *et al.* Experimental evidence for perovskite and post-perovskite coexistence throughout the whole D'' region. *Earth Planet. Sci. Lett.* **293,** 90–96 (2010).
25. Murakami, M., Hirose, K., Sata, N. & Ohishi, Y. Post-perovskite phase transition and mineral chemistry in the pyrolitic lowermost mantle. *Geophys. Res. Lett.* **32,** L03304 (2005).
26. Davaille, A. A simultaneous generation of hotspots and superswells by convection in a heterogeneous planetary mantle. *Nature* **402,** 756–760 (1999).
27. Labrosse, S., Hernlund, J. W. & Coltice, N. A crystallizing dense magma ocean at the base of the Earth's mantle. *Nature* **450,** 866–869 (2007).
28. Hirose, K., Shimizu, N., van Westrenen, W. & Fei, Y. Trace element partitioning in Earth's lower mantle and implications for geochemical consequences of partial melting at the core-mantle boundary. *Phys. Earth Planet. Inter.* **146,** 249–260 (2004).
29. Corgne, A., Liebske, C., Wood, B. J., Rubie, D. C. & Frost, D. J. Silicate perovskite-melt partitioning of trace elements and geochemical signature of a deep perovskitic reservoir. *Geochim. Cosmochim. Acta* **69,** 485–496 (2005).

30. Liebske, C., Corgne, A., Frost, D. J., Rubie, D. C. & Wood, B. J. Compositional effects on element partitioning between Mg-silicate perovskite and silicate melts. *Contrib. Mineral. Petrol.* **149,** 113–128 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank N. Bolfan-Casanova, M. A. Bouhifd, T. Druitt, T. Hammouda and J.-M. Hénot for help and discussions. This work is supported by the French National Centre for Scientific Research's National Institute for Earth Sciences and Astronomy, the ESRF and the European C2C programme. This is Laboratory of Excellence ClerVolc contribution no. 26.

Author Contributions D.A., S.P., G.L.N., G.G. and M.M. synthesized the sample and took the XRD and XRF measurements at the ID27 beamline. S.P. and G.V. took the XRF measurements at the ID21 beamline. D.A. and J.-L.D. performed the electron-probe micro-analyses at the Laboratoire Magmas et Volcans. D.A., S.P. and G.L.N. performed the data treatment and wrote the paper. All authors discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.A. (d.andrault@opgc.univ-bpclermont.fr).

METHODS

Sample preparation. We used a synthetic (Ca,Mg,Al,Si,Fe) glass with oxide contents in chondritic proportions, except for iron with a Fe/(Σ cations) ratio of 6.1. This represents a model composition for the primitive mantle after core segregation (similar to pyrolite, and the same sample as in ref. 10). We did not include minor and trace elements, among which the most abundant are Na (4,900 p.p.m.) and K (560 p.p.m.), because their effect on Fe partitioning can be neglected to a first approximation.

We performed experiments at high pressure and high temperature using the laser-heated diamond anvil cell (LH-DAC) technique. We used diamond anvils with flat culets of 250-μm diameter or bevel-type culets of 100–300-μm diameter. Very thin (5–10 μm) samples were loaded in tungsten or rhenium gaskets between two NaCl pellets of 5–10-μm thickness each. After we compressed the sample to the target pressure, we heated the samples using two Nd-doped Y-Al-garnet or Y-fibre lasers with a laser-spot size of more than 20-μm diameter. The temperature was measured at the centre of the laser hotspot by analysing the pyrometric signal emitted over a sample area of $3 \times 3\ \mu m^2$. Melting of samples throughout their entire thickness was induced by increasing laser power until the liquidus temperature was reached at the centre of the sample. Melting criteria are based on the use of *in situ* X-ray diffraction; details are reported elsewhere[10]. The laser power was maintained for between a couple of seconds (for six samples) and about 1 min (for three samples) before rapid quenching. The use of two different heating procedures was dictated by the need to assess the significance of the artefact that chemical diffusion could produce in our samples. X-ray diffraction patterns show major NaCl contribution (Fig. 2), a sign of good thermal insulation between the laser-heated sample and the diamonds. This minimizes the axial temperature gradient in the sample. No chemical reactions between NaCl and our samples could be detected on the basis of X-ray diffraction or electron-probe microanalyses.

The NaCl equation of state was used to derive the pressure at room temperature[31], before and after laser heating. In this paper, we report experimental runs at nine different nominal pressures. We estimate the pressure correction (ΔP) associated with the laser heating in a partially isochoric regime using the same technique as in a previous study[10]. This yields ΔP (GPa) $= 2.5 \times 10^{-3}\ \Delta T$(K), where $\Delta T = T - 300$ (ΔT, change in temperature; T, temperature of sample synthesis). On the basis of nominal pressures and experimental temperatures reported in Table 1, we recalculated the pressure encountered by each sample at high temperatures.

X-ray diffraction. X-ray diffraction was performed at the ID27 beamline of the ESRF, simultaneously with a first set of XRF measurements (see below). Two under-vacuum undulators generate an X-ray flux of 8.4×10^{11} photons s^{-1} for an energy tuned to a wavelength of 0.3738 Å. The X-ray beam was focused by two Kirkpatrick–Baez mirrors to less than 2 μm × 2 μm full-width half-maximum spot on the sample. We used the MAR-160 charge-coupled-device detector with a typical acquisition time of 20–30 s. Diffraction images were analysed and integrated using the Fit2d program[32]. We refined phase contents by performing multiphase Rietveld refinements using the General Structure Analysis System code[33]. According to the composition of the starting material, proportions of Mg-Pv and Ca-bearing (Ca-Pv) perovskite phases and Fp are expected to be 75.7, 4.5 and 19.8 mol%, respectively. We note a weak Ca-Pv signal for some of the samples, because the Ca-Pv amorphizes partially on decompression and because rapid quenching from the molten state could lead to other Ca-rich polymorphs, such as the "new aluminous phase"[34]. We also note a weak precision for the determination of Fp contents, because its major diffraction peaks overlap with those of the major Mg-Pv phase.

X-ray fluorescence. The experimental set-up of ID27 allows X-ray fluorescence (XRF) measurements *in situ* in the LH-DAC[35]. However, we chose to take the measurements after release of pressure and temperature, to prevent the absorption of the emitted Fe fluorescence signal by the diamond window. Similar measurements have been performed in the DAC, even for elements lighter than Fe. For example, dissolution of TiO_2 in geological fluids has been addressed *in situ* at high pressure and temperature, for concentrations as low as a few parts per million[36]. In our case, the Fe content is a few orders of magnitude higher in terms of concentration than in many other studies dealing with liquid solutions (such as ref. 37), and does not specifically need to be measured by a low-energy beam. The originality of our approach resides in the extreme conditions needed for the sample synthesis, resulting in very small regions of interest, of a few $10\ \mu m^2$.

At the ID27 beamline, we used an energy-dispersive VorteX (SII NanoTechnology USA Inc.) Silicon Drift Detector in two different configurations. For the first set of experiments, we set the detector behind the sample at about 60° from the incident beam. It is the best angle that can be achieved in the transmission mode when using non-transparent gaskets. The detector was protected from incoherent X-ray signal using an Ag collimator. To maximize the photoelectric effects and minimize Compton and Rayleigh scattering, we performed a second set of experiments in backscattering mode with the detector set at about 90° from the incoming beam. As a collimator, we used a polycapillary X-ray half-lens with the role of enhancing the XRF signal-to-noise ratio[38]. For both set-ups, the samples were rotated by 8° towards the detector to optimize the angle of XRF emission available for XRF measurements. The sample-to-detector distance was about 20 mm. We detected K-lines of Ca and Fe from the sample, K-lines of Cl from the NaCl pressure medium, L-lines of W or Re from the gasket material, and L-lines of the Pb used as the X-ray absorber on the beamline. Typical acquisition time was 100 or 150 s per spectrum.

We also performed high-resolution XRF mapping at the ID21 beamline. We used a Si <111> double crystal monochromator to tune the incoming X-ray energy to 7.2 keV. This energy is optimized for the photo-excitation of 1*s* electrons from Fe (binding energy of 7.112 keV) followed by K_α and K_β fluorescence emission. This results in a less-efficient XRF count for Mg and Si, because their absorption cross sections are largest just above their 1*s* binding energies (1.305 keV and 1.839 keV, respectively), and decrease as the excitation energy increases. Nevertheless, the use of a high-vacuum (around 0.001-Pa) sample environment allowed the successful detection of the fluorescence lines of Al, Si, Cl, Ca and Fe, for an exposure time of 1 s per pixel. The monochromatic beam was focused down to 0.2 μm × 0.7 μm using a tungsten Fresnel Zone Plate lens, resulting in a photon flux of about 5×10^9 photons s^{-1}. The scanning X-ray microscope is housed in an environmental chamber allowing operation in vacuum (0.01–0.0001 Pa). The sample is rastered in the X-ray focal plane using a combination of mechanical stages and a piezo-driven monolithic XY flexure stage. The system is equipped with integrated capacitive encoders allowing positioning resolutions of about 10 nm. XRF detection was achieved with a large-active-area (80-mm^2) detector (XFlash 5100 Bruker Silicon Drift Diode), the distance to the sample of which was adjusted to keep the dead time below 20%. Using this technique, we obtained very-high-resolution fluorescence maps for Fe content.

Calculation of D_{Fe} and its associated uncertainties. In the XRD and XRF maps, we identify three different sample regions with Mg-Pv liquidus phase, quenched liquid and untransformed starting material. Characteristic XRF intensity of each type of material is calculated by averaging XRF signals measured over typical sample surfaces (Supplementary Fig. 1). The Fe content for the liquid phase is obtained by integrating XRF intensities over the whole sphere-shell structure. Then $D_{Fe} = [Fe]_{Solid}/[Fe]_{Liquid}$ is calculated from the ratio of XRF intensities measured in juxtaposed Mg-Pv and quenched-liquid phases. The unheated starting material can also be used as an internal standard (with Fe/(Σ cations) = 6.1%) to transform XRF intensities into atomic percentages. Similar calculations have been previously used for high-pressure samples[39,40]. For the sample synthesized at 105 GPa and 4,150 K (Supplementary Fig. 3), for example, variation in XRF intensity between the different sample regions implies Fe content of 4.0% and 7.5% in Mg-Pv and quenched-liquid phases, respectively. It yields a D_{Fe} value of 0.54(3).

There are different sources of uncertainties for D_{Fe}. The first, of the order of 0.1% (Supplementary Table 1), originates from the fit to the experimental XRF profiles used to retrieve the XRF intensity of Fe at each sample position. For the second, based on the knowledge of the XRF intensity emitted by Fe at each pixel, we obtain Fe maps for all the samples considered in this study (Supplementary Fig. 1). Superimposed on each Fe map, we draw sample areas, which include between 13 and 298 pixels for the Pv region, and between 4 and 72 pixels for the liquid regions. We average the XRF signal over these sample areas and retrieve mean values and standard deviations for XRF intensities emitted by Fe in both liquid and perovkite regions (Supplementary Table 1). In the third source of uncertainty, D_{Fe} values are retrieved from the ratio between Fe-XRF intensities measured in the perovskite and liquid sample regions. This calculation neglects the difference in matrix corrections that should be done for the two different sample regions. The assumption seems mostly reasonable, because Fe is the heaviest element in the sample (and thus its XRF radiation is poorly reabsorbed by the sample) and because both silicate phases have comparable electronic densities. Further experimental uncertainty could also arise from a possible variation of sample thickness as a function of sample position. In a previous report[3] of a similar experiment performed using an Ar pressure medium (instead of NaCl, as in our study), it was shown that the sample thickness is basically unchanged after laser heating. Our observations are also compatible with an atomic diffusion restricted to small distances. This is particularly true for the six samples heated for only a few seconds, for which the shape of the original piece of glass remained unchanged (Fig. 1a). However, thickness variations up to 10% could be below our detection limit. We therefore use this value as an uncertainty. Combining all the sources of uncertainty described above, we estimate the total uncertainty on D_{Fe} partition coefficients to be between 0.09 and 0.13, for D_{Fe} values between 0.47 and 0.6 (Table 1).

31. Sata, N., Shen, G., Rivers, M. L. & Sutton, S. R. Pressure–volume equation of state of the high-pressure *B2* phase of NaCl. *Phys. Rev. B* **65,** 114114–114117 (2002).
32. Hammersley, J. *Fit2d User Manual* (ESRF, 1996).

33. Larson, A. C. & Von Dreele, R. B. *GSAS Manual* (Los Alamos National Laboratory, 1988).
34. Miyajima, N., Fujino, K., Funamori, N., Kondo, T. & Yagi, T. Garnet–perovskite transformation under conditions of the Earth's lower mantle: an analytical transmission electron microscopy study. *Phys. Earth Planet. Inter.* **116,** 117–131 (1999).
35. Petitgirard, S. *et al.* An *in situ* approach to study trace element partitioning in the laser heated diamond anvil cell. *Rev. Sci. Instrum.* **83,** 013904 (2012).
36. Manning, C. E., Wilke, M., Schmidt, C. & Cauzid, J. Rutile solubility in albite-H_2O and $Na_2Si_3O_7$-H_2O at high temperatures and pressures by in-situ synchrotron radiation micro-XRF. *Earth Planet. Sci. Lett.* **272,** 730–737 (2008).
37. Mayanovic, R. A., Yan, H., Anderson, A. J., Meredith, P. R. & Bassett, W. A. In situ X-ray absorption spectroscopic study of the adsorption of Ni^{2+} on Fe_3O_4 nanoparticles in supercritical aqueous fluids. *J. Phys. Chem. C* **116,** 2218–2225 (2012).
38. Wilke, M. *et al.* A confocal set-up for micro-XRF and XAFS experiments using diamond-anvil cells. *J. Synchrotron Radiat.* **17,** 669–675 (2010).
39. Sanchez-Valle, C. *et al.* Dissolution of strontianite at high *P*–*T* conditions: an in-situ synchrotron X-ray fluorescence study. *Am. Mineral.* **88,** 978–985 (2003).
40. Petitgirard, S. *et al.* A diamond anvil cell for X-ray fluorescence measurements of trace elements in fluids at high pressure and high temperature. *Rev. Sci. Instrum.* **80,** 033906 (2009).

LETTER

doi:10.1038/nature11264

Seasonal bone growth and physiology in endotherms shed light on dinosaur physiology

Meike Köhler[1], Nekane Marín-Moratalla[2], Xavier Jordana[2] & Ronny Aanes[3,4]

Cyclical growth leaves marks in bone tissue that are in the forefront of discussions about physiologies of extinct vertebrates[1]. Ectotherms show pronounced annual cycles of growth arrest that correlate with a decrease in body temperature and metabolic rate; endotherms are assumed to grow continuously until they attain maturity because of their constant high body temperature and sustained metabolic rate[1,2]. This apparent dichotomy has driven the argument that zonal bone denotes ectotherm-like physiologies, thus fuelling the controversy on dinosaur thermophysiology and the evolution of endothermy in birds and mammal-like reptiles[1–4]. Here we show, from a comprehensive global study of wild ruminants from tropical to polar environments, that cyclical growth is a universal trait of homoeothermic endotherms. Growth is arrested during the unfavourable season concurrently with decreases in body temperature, metabolic rate and bone-growth-mediating plasma insulin-like growth factor-1 levels, forming part of a plesiomorphic thermometabolic strategy for energy conservation. Conversely, bouts of intense tissue growth coincide with peak metabolic rates and correlated hormonal changes at the beginning of the favourable season, indicating an increased efficiency in acquiring and using seasonal resources. Our study supplies the strongest evidence so far that homeothermic endotherms arrest growth seasonally, which precludes the use of lines of arrested growth as an argument in support of ectothermy. However, high growth rates are a distinctive trait of mammals, suggesting the capacity for endogenous heat generation. The ruminant annual cycle provides an extant model on which to base inferences regarding the thermophysiology of dinosaurs and other extinct taxa.

Inferences about the thermophysiology of extinct vertebrates are largely based on bone histology. This approach rests on the premise that high rates of bone formation are possible only in systems with elevated metabolic rates[3]. Ectotherms, which almost exclusively use external heat sources for their metabolic processes, show annual cycles of alternating zones of slowly formed lamellar bone and lines of arrested growth (LAGs). Conversely, endotherms (mammals and birds) generate heat internally that entails high and constant rates of metabolism and high growth rates. Accordingly, endotherm bone matrix is assumed to consist of continuously growing, well-vascularized fibrolamellar tissue without LAGs[1,2]. Cyclical bone growth has therefore been associated with ectothermic physiology[1]. Indeed, in a seminal book[5] on mammalian hard tissues it is explicitly stated that LAGs do not form in mammalian fibrolamellar bone. This notion still persists in current literature[1,2]. Increasing sampling, however, showed that the vast majority of dinosaurs possessed a unique and contradictory kind of

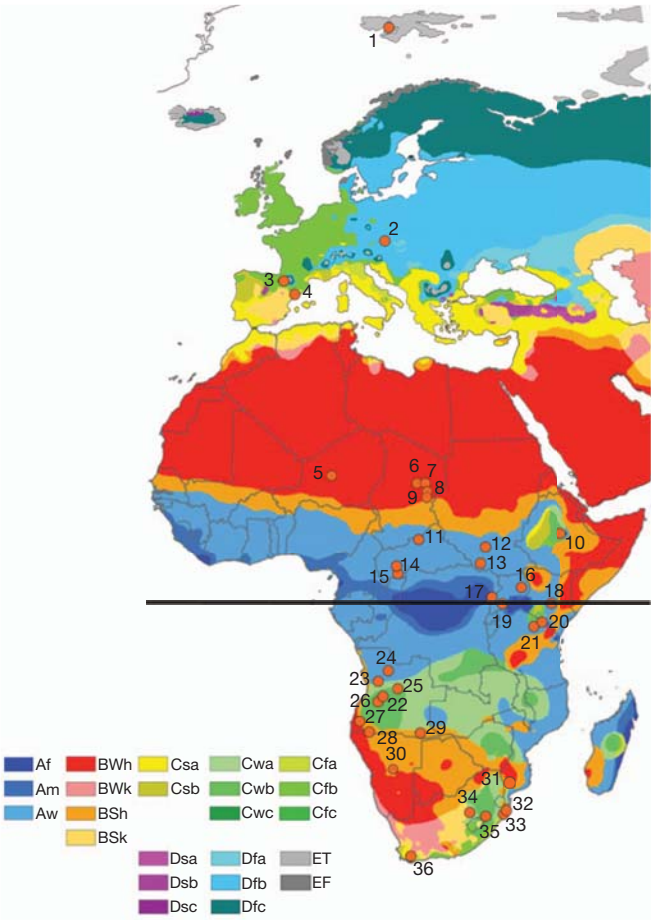


Figure 1 | World map of Köppen-Geiger climate classification[30], showing the climate zones of sample sites. Northern Hemisphere: ET, Polar, tundra, no true summer (Svalbard (1)); Dfb, Humid Continental, mild summer, long cold winter (Vienna (2)); Cfb, Oceanic, no dry season, warm summer (Huesca (3)); Csa, Temperate Mediterranean, dry and hot summer (Alfarc (4)); BWh, Desert, precipitation less than half of potential evapotranspiration, average temperature more than 18 °C (Agadez (5), Ouadi Achim (7), Biltine (8)); BSh, Semiarid Steppe, annual precipitation greater than BWh climates (Oum Chaluba (6), Abéché (9)); Aw, Tropical Wet and Dry Savanna, pronounced dry winter, constant high temperatures (Tiri (11), Tonj (12), Yambio (13), Mbata (14), Bakota (15), Komolo (16), Cunni (10)); Am, Monsoon (Semliki (17)). Southern Hemisphere: Aw, Tropical Wet and Dry Savanna (Nyeri (18), Lake Mburo (19), Camata (24), Tumba grande (23), Lolkisale (21)); BWh, Desert (Capolopopo (27)); BSh, Semiarid Steppe (Ruacaná (28), Dirico (29), Ojitsewa (30), Massingir (31)); BSk, Middle Latitude Semiarid Steppe (Willem Pretorius Park (34), Bredasdorp (36)); Csb, Temperate Mediterranean, dry summer subtropical (Longido (20)); Cwb, Oceanic, changeable weather, rainy cool summer, dry noticeable winter (Sanguengue (22), Cambembe (25), Culele (26), Giants Castle (35)); Cfa, Humid Subtropical, warm–hot summer, dry cold winter (False Bay (32), Umfolozi (33)). Main climates: A, equatorial; B, arid; C, warm temperate; D, cold continental; E, polar. Precipitation: W, desert; S, steppe; f, fully humid; s, summer dry; w, winter dry, m, monsoonal. Temperature: h, hot arid, low latitude; k, cold arid, middle latitude; a, hot summer; b, warm summer; c, cold summer.

[1]ICREA at the Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain. [2]Institut Català de Paleontologia Miquel Crusafont, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain. [3]Norwegian Polar Institute, Fram Centre, NO-9296 Tromsø, Norway. [4]Norwegian Directorate for Nature Management, NO-7047 Trondheim, Norway.

bone tissue, with zones of rapidly deposited fibrolamellar bone separated by LAGs[6]. This combination indicated fast, yet interrupted, growth from year to year. LAGs are also increasingly found in mammals, although their significance remains unclear[7]. This prompted opposing hypotheses[2,4] and conjectures[8] concerning the physiological and ecological correlates of zonal bone. However, much of the debate stems from the lack of a detailed, methodical study on extant endotherms on which to base inferences regarding extinct vertebrates[7]. For this reason we used an exceptionally large and complete sample of wild ruminants to explore the association of bone growth cycles, including rest line formation, with seasonal rainfall and temperature patterns, annual cycles of body core temperature, resting metabolic rate and other physiological variables. This study provides insights into physiological and endocrine changes associated with annual dynamics of zonal bone growth, and a link between bone tissue type and seasonal metabolic physiology. Finally, it provides the foundation for discussions on the physiology of extinct taxa that has been lacking for the past 40 years.

Ruminants are, in many respects, an excellent group on which to conduct such a study: first, they are a geologically young group of modern large endotherms; second, they represent the extreme state of homeothermic endothermy as a result of the thermal needs of their complex four-chambered stomach[9]; third, most species have a relatively long juvenile period with osteogenesis continuing beyond at least one year, a basic condition for the formation of fibrolamellar-zonal bone; and fourth, ruminants dwell in a wide array of climatic regimes, allowing the identification of the ecological correlates of rest lines.

Our ruminant sample comprises more than 100 wild individuals belonging to almost all ruminant tribes (Supplementary Table 1, Supplementary Fig. 1 and Supplementary Methods). Many of these species are protected nowadays, and skeletal material is extremely scarce and difficult to obtain. They come from 36 African and European localities under almost all Köppen–Geiger climate regimes (Fig. 1). This ecological diversity is important in this study because net primary productivity of biomes is highly sensitive to annual precipitation and is therefore a principal constraint on growth and physiology in mammals[10].

Associated data on body mass, latitude, climate zone, and date of death (Supplementary Table 1) cover the ecological aspects, and data on thermometabolic physiology and hormone levels, monitored for two populations of our sample (Supplementary Methods), cover the physiological aspects. Seasonal growth patterns were reconstructed from bone histology (Supplementary Methods).

LAGs are universally present in all climatic regimes, from high and cold to low and hot latitudes, and from moist tropical forests to arid

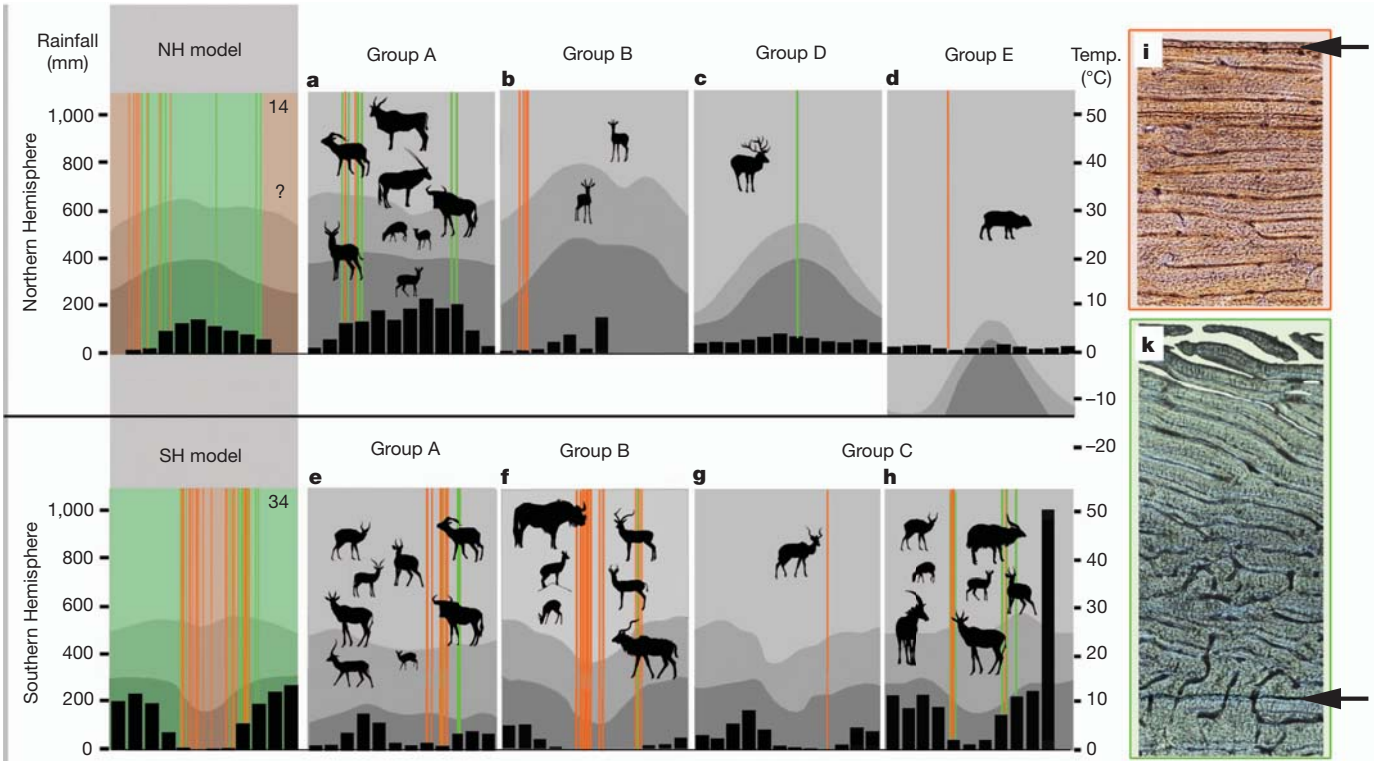


Figure 2 | Climatic context of rest line formation. Selection of African and European localities covering the principal Köppen–Geiger climate zones, organized into Northern Hemisphere (**a**–**d**) and Southern Hemisphere (**e**–**h**). Average monthly rainfall (black bars; each bar represents a month starting with January) and temperature (light curve, maximum; dark curve, minimum) are contrasted with growth stage at death in individuals dwelling in these localities (orange bars and **i**, growth arrest; green bars and **k**, active growth; each bar represents one individual; see Supplementary Methods). Some neighbouring localities were combined when seasonal rainfall and temperature patterns were comparable. The NH and SH models summarize the patterns for the Northern Hemisphere and the Southern Hemisphere, respectively (sample size indicated at upper right). The emerging pattern of seasonality of growth and growth arrest is compelling despite a data gap (question mark in NH) caused by the timing of hunting seasons. Both models show that the period of growth arrest is limited to a 3–4-month (orange) period of drought or low precipitation. Mixed information at the limit between growth and growth arrest probably results from minor climatic differences between sites and between years of death. Sites representing climate zones for northern localities: **a**, Tropical (Aw, Tropical Savanna; Am, Monsoon), Central African Republic (Bakota); **b**, Arid (BW, Desert; BS, Steppe), Niger (Agadez); **c**, Continental (Dfb, no dry season), Austria (Vienna); **d**, Polar (ET, Tundra), Norway (Svalbard). Sites representing climate zones for southern localities: **e**, Tropical (Aw, Tropical Savanna), Kenya (Nyeri); **f**, Arid (BW, Desert; BS, Steppe), Angola (Ruacaná); **g**, Temperate (Csb, Mediterranean), Tanzania (Longido); **h**, Temperate (Cfa, Cwb, dry winter, rainy summer), Angola (Huambo). **i**, Examples for arrested growth (*Cephalophus callipygus*, Mbata ICP 56243). **k**, Example for active growth (*Cervus elaphus*, Vienna ICP 56308). Arrows indicate last LAG. Individuals used in this analysis are marked grey in Supplementary Table 1.

deserts. They form annually in the fast-growing, highly vascularized fibrolamellar bone tissue of ruminants of all body sizes ranging from the 3.2-kg pygmy antelope (*Nesotragus moschatus*) to the 900-kg giant eland (*Tragelaphus derbianus*), resulting in a bone tissue pattern typically found in all dinosaurs except sauropods[11]. All species show the same dynamics of cyclical bone growth, whereby transitions between principal tissue types indicate important annual changes in growth rate (Supplementary Fig. 2).

In both the Northern and Southern Hemispheres, LAGs form during the energetically challenging, usually dry (at high latitudes cold), season (Fig. 2). This is in accordance with the notion that precipitation-mediated resource availability influences growth and physiology[10]. Both hemispheres show inverted seasonal patterns of growth and growth arrest in accordance with the inverted patterns of precipitation (dry regions are centred near 20° N during January to April, and near 15° S during June to September[12]). Thus, north of the Equator (Fig. 2a–c) rest lines form from February to April (data for January are not available; see the legend to Fig. 2); south of the Equator (Fig. 2e–h) rest lines form from May to September. At higher northern latitudes and with increasingly longer winters, however, growth is arrested until April or May (Fig. 2d).

Transitions between the principal bone tissues[13] also follow a seasonal pattern: they form part of the same annual cycle. Fibrolamellar tissue with the highest degree of vascularization (circumferential or reticular organization) is deposited during the zenith of the favourable rainy (warm at high latitudes) season.

Ruminants, in common with all organisms exposed to seasonal environments, have evolved physiological adaptations that allow them to cope with unfavourable seasons but also to exploit the favourable seasons. Data monitored for two species of our sample illustrate the annual cycle of endocrine changes (Svalbard reindeer, *Rangifer tarandus platyrhynchus*) and of thermophysiological changes (alpine red deer, *Cervus elaphus*).

In the Svalbard population, LAGs form during the polar winter (Fig. 2d) coinciding with the lowest voluntary food intake[9], depletion of body fat reserves[14] and high rates of mortality[15]. During this period of low energy budget, reindeer show high levels of growth hormones (responsible for the mobilization of stored fat[16]), low levels of thyroxine (responsible for acceleration of the metabolic rate[16]) and correspondingly low resting metabolic rate[9]. Plasma insulin-like growth factor-1 (IGF-1) levels, which have a critical function in the development of the growing skeleton by establishing both longitudinal and transverse bone accrual[17], reach a nadir in February (continental subspecies *R. t. tarandus*[18]), well within the period of rest line formation in our *R. t. platyrhynchus* sample. Growth arrest is therefore concurrent with hormonal and metabolic nadirs.

Annual cycles of heart rate (a proxy for metabolic rate) and rumen temperature (a proxy for body core temperature) have been monitored in our alpine deer[19]. During the winter season, red deer resorts to the use of hypometabolism by decreasing the setpoint of body core temperature to decrease thermometabolic costs[19,20]. The first LAG in the bone tissue of a male yearling records this thermophysiological state (Fig. 2k). The winter LAG is followed by fibrolamellar tissue with progressively increasing vascular spaces until July (the month of death; see Supplementary Methods). This bout of intense growth coincides with the highest annual levels of heart rate (May to July), rumen temperature (June to July), body mass (June to July) and food consumption (July to September) reported for this population[19], confirming the tight scheduling of maximum growth rate and maximum metabolic rate with times of highest energy availability. Concordant with these data, the highest levels of bone-growth-promoting plasma IGF-1 in arctic ruminants (wild *Ovibos moschatus* and *Rangifer tarandus*) are attained between July and August[21] and between August and October[18], respectively.

The existence of annual cycles of simultaneous physiological and endocrine changes in ruminants is supported by scattered data on a large number of other species analysed here. These can be summarized as follows. During the unfavourable (usually dry) season, ruminants use a complex energy-conserving strategy that includes cyclical variation in body core temperature, resting metabolic rate, and associated changes in hormonal levels[9,16,18–27]. Growth arrest forms part of this strategy. Conversely, at the beginning of the favourable (usually humid) season, when resources with high nutritive values become plentiful, ruminants maximize growth rate (associated with enhanced thyroid activity[27] and IGF-1 increase[18]) and raise their metabolic rate. Peak metabolic and IGF-1 levels coincide with bouts of intense growth in the middle of the favourable season[23].

The annual cycle of ruminants provides a model for the correlation between cyclical bone growth and seasonal physiology (Fig. 3).

Seasonal growth cycles have been reported for almost all ectotherms and are considered to be a plesiomorphy. The presence of such cycles in modern homeothermic endotherms and their fine-tuning with seasonal fluctuations in resource availability suggest that they form part of an ancient inheritance that is still operative today. In the annual cycle of endotherms the capacity to cease growth during periods of negative energy budgets is not new. What appears to be new, however, is the pervasive association of maximum rates of bone accrual with high, sustained metabolic rates during the favourable season. This clearly apomorphic feature reflects the capability of efficiently exploiting and allocating abundant resources to growth when food is plentiful.

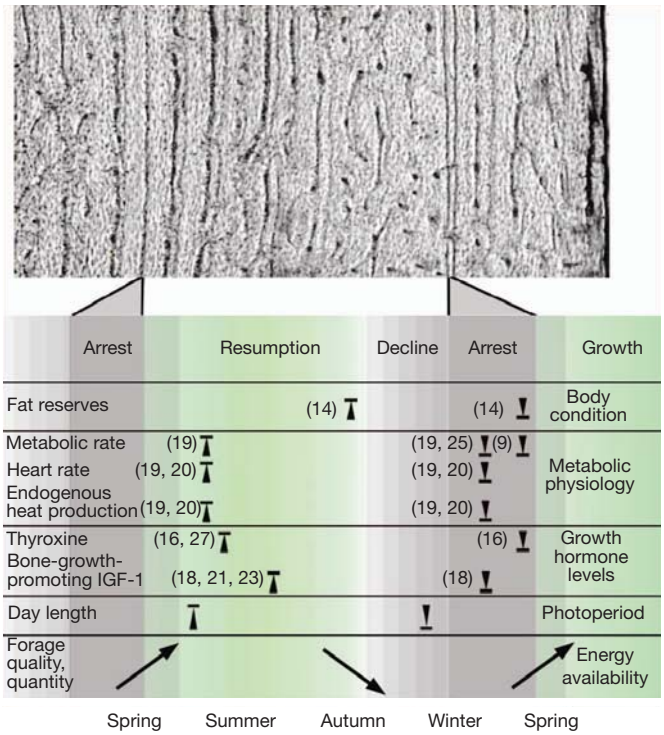


Figure 3 | A ruminant model for the correlation between bone tissue cycles and seasonal physiology. Synchronization of growth, hormone levels and thermophysiology with environmental conditions (energy availability and photoperiod) over an annual cycle. The combined effect increases reproductive fitness by mitigating the effects of negative energy balance on body condition during the unfavourable season, and by improving body condition during the favourable season. The annual histological cycle (top) for *Aepyceros melampus* (ICP 56215) is tentatively adapted to the seasonal physiological cycle (bottom). LAGs comprise a period of arrested growth, here schematically represented by a broad grey bar; lighter shadowed bars indicate growth decline and resumption. Upward directed arrowheads, annual zenith of the trait; downward directed arrowheads, annual nadir of the trait. In this model the schedule of changes in physiological and hormonal traits comes from Northern Hemisphere higher-latitude ruminants because these have been studied more extensively (references in parentheses).

This capability results in high growth rates that contrast with the low growth rates of ectotherms (only 1/10 to 1/30 of those of endotherms of the same body mass[28]), and is in accordance with the notion of metabolic constraints on maximum growth rate[29].

The consistently seasonal formation of rest lines in homeothermic endotherms debunks the key argument from bone histology in support of dinosaur ectothermy. Our study instead suggests that the extensive vascularization of the fibrolamellar bone in most dinosaurs and other extinct vertebrates is tightly correlated with seasonal maxima of endogenous heat production, an association that should be explored in future studies.

METHODS SUMMARY

We used 115 right femora of extant ruminants. We cut segments roughly 2 cm in length from the central part of the diaphyses. The samples were embedded in epoxy resin (Araldite 2020). The surface of interest was exposed with a Buehler Isomet low-speed saw, and later polished on a glass sheet coated with carborundum powder, in decreasing particle size (for example 600, 800 and 1,000 grit). Each sample was fixed to a frosted glass slice with ultraviolet curing glue (Loctite 358). Ground sections were prepared with a diamond saw (PetroThin; Buehler) to a final thickness of about 100–120 μm. The thin sections were polished with a gradient of carborundum (800 and 1,200 grit), dehydrated through a graded series of alcohol baths, cleared in Histo-Clear II for 5 min and finally mounted in DPX mounting medium. Once finished, the slides were observed under circularly polarized transmitted light with a 1λ filter, and micrographs were taken with a Leica camera under a Leica DM 2500P polarization microscope.

Received 17 May; accepted 25 May 2012.
Published online 27 June 2012.

1. Fastovsky, D. E. & Weishampel, D. B. *Dinosaurs. A Concise Natural History* (Cambridge Univ. Press, 2009).
2. Chinsamy, A. & Hillenius, J. in *The Dinosauria* 2nd edn (eds Weishampel, D., Dodson, B. P. & Osmolska, H.) 643–659 (Univ. of California Press, 2004).
3. Bennett, A. F. & Ruben, J. A. in *The Ecology and Biology of Mammal-like Reptiles* (eds Hotton, N., MacLean, P. D., Roth, J. J. & Roth, E. C.) 207–218 (Smithsonian Institution Press, 1986).
4. Padian, K. & Horner, J. R. in *The Dinosauria* 2nd edn (eds Weishampel, D., Dodson, B. P. & Osmolska, H.) 660–671 (Univ. of California Press, 2004).
5. Klevezal, G. A. *Recording Structures of Mammals. Determination of Age and Reconstruction of Life History* (A.A. Balkema, 1996).
6. Reid, R. H. Primary bone and dinosaurian physiology. *Geol. Mag.* **121,** 589–598 (1984).
7. Sander, P. M. & Andrássy, P. Lines of arrested growth and long bone histology in Pleistocene large mammals from Germany: what do they tell us about dinosaur physiology? *Palaeontographica A* **277,** 143–159 (2006).
8. Chinsamy, A., Thomas, D. B., Tumarkin-Dertzian, A. R. & Fiorillo, A. R. Hadrosaurs were perennial polar residents. *Anat. Rec.* **295,** 610–614 (2012).
9. Tyler, N. J. C. & Blix, A. S. Survival strategies in arctic ungulates. *Rangifer* **3,** 211–230 (1990).
10. McNab, B. Geographic and temporal correlations of mammalian size reconsidered: a resource rule. *Oecologia* **164,** 13–23 (2010).
11. Klein, N. & Sander, M. Ontogenetic stages in the long bone histology of sauropod dinosaurs. *Paleobiology* **34,** 247–263 (2008).
12. Gruber, A. & Levizzani, V. *Assessment of Global Precipitation Products. A Project of the World Climate Research Programme Global Energy and Water Cycle Experiment (GEWEX) Radiation Panel* (WCRP Report no. 128, WMO/TD no. 1430) (World Meteorological Organization, 2008).
13. de Margerie, E., Cubo, J. & Castanet, J. Bone typology and growth rate: testing and quantifying 'Amprino's rule' in the mallard (*Anas platyrhynchus*). *C. R. Biol.* **325,** 221–230 (2002).
14. Reimers, E. Body composition and population regulation of Svalbard reindeer. *Rangifer* **4,** 16–21 (1984).
15. Aanes, R., Sæther, B.-E. & Øritsland, N. A. Fluctuations of an introduced population of Svalbard reindeer: the effects of density dependence and climatic variation. *Ecography* **23,** 437–443 (2000).
16. Ringberg, T. The Spitzbergen reindeer—a winter-dormant ungulate? *Acta Physiol. Scand.* **105,** 268–273 (1979).
17. Courtland, H.-W. *et al.* Serum IGF-1 affects skeletal acquisition in a temporal and compartment-specific manner. *PLoS ONE* **6,** e14762 (2011).
18. Bubenik, G. A. *et al.* Seasonal levels of metabolic hormones and substrates in male and female reindeer (*Rangifer tarandus*). *Comp. Biochem. Physiol.* **120,** 307–315 (1998).
19. Turbill, C., Ruf, T., Mang, T. & Arnold, W. Regulation of heart rate and rumen temperature in red deer: effects of season and food intake. *J. Exp. Biol.* **214,** 963–970 (2010).
20. Arnold, W. *et al.* Nocturnal hypometabolism as an overwintering strategy of red deer (*Cervus elaphus*). *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **286,** R174–R181 (2004).
21. Lawler, J. P. & White, R. G. Seasonal changes in metabolic rates in muskoxen following twenty-four hours of starvation. *Rangifer* **17,** 135–138 (1997).
22. Piccione, G., Giannetto, C., Casella, S. & Caola, G. Annual rhythms of some physiological parameters in *Ovis aries* and *Capra hircus*. *Biol. Rhythm Res.* **40,** 455–464 (2009).
23. Suttie, J. M. & Webster, J. R. Extreme seasonal growth in arctic deer: comparisons and control mechanisms. *Am. Zool.* **35,** 215–221 (1995).
24. Hetem, R. S. *et al.* Variation in the daily rhythm of body temperature of free-living Arabian oryx (*Oryx leucoryx*): does water limitation drive heterothermy? *J. Comp. Physiol.* **180,** 1111–1119 (2010).
25. Signer, C., Ruf, T. & Arnold, W. Hypometabolism and basking: the strategies of Alpine ibex to endure harsh over-wintering conditions. *Funct. Ecol.* **25,** 537–547 (2011).
26. Ostrowski, S., Mesochina, P. & Williams, J. B. Physiological adjustments of Sand Gazelles (*Gazella subgutturosa*) to a boom-or-bust economy: standard fasting metabolic rate, total evaporative water loss, and changes in the sizes of organs during food and water restriction. *Physiol. Biochem. Zool.* **79,** 810–819 (2006).
27. Todini, L. Thyroid hormones in small ruminants: effects of endogenous, environmental and nutritional factors. *Animal* **1,** 997–1008 (2007).
28. McNab, B. K. *The Physiological Ecology of Vertebrates. A View from Energetics* (Cornell Univ. Press, 2002).
29. Montes, L., Castanet, J. & Cubo, J. Relationship between bone growth rate and bone tissue organization in amniotes: first test of Amprino's rule in a phylogenetic context. *Animal Biol.* **60,** 25–41 (2010).
30. Peel, M. C., Finlayson, B. L. & McMahon, T. A. Updated world map of the Köppen–Geiger climate classification. *Hydrol. Earth Syst. Sci. Discuss.* **11,** 1633–1644 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Th. Kaiser for permission to cut femora of skeletons from the Oboussier collections and from zoological material housed at the Zoological Institute and Museum of the University of Hamburg; W. Arnold for providing alpine red deer material, and A. Kübber for preparing and sending it; R. García González for providing red deer femora from Jaca (Spanish Pyrenees) and *Capra ibex* from Alfarc (Tarragona, Spain); all the people that helped collect the Svalbard material, and R. García for preparation of the thin sections; and J. Horner, H. Woodward, S. Moyà-Solà, T. Bromage and J. Cubo for comments on the manuscript. This work was supported by the Spanish Ministry of Science and Innovation (CGL2008-06204/BTE, 2012: CGL2011-24685, M.K.; BES-2009-02641, N.M.-M.; JCI-2010-08157, X.J.); the work was partly funded by the Norwegian Research Council (NORKLIMA 178561/S30, R.A.). The material is tabulated in the Supporting Online Material and archived at the Institut Català de Paleontologia, Catalonia, Spain.

Author Contributions: M.K. designed the study and wrote the manuscript. R.A. gathered the Svalbard material and was involved in discussions about the biology of Svalbard reindeer. M.K., N.M.-M. and X.J. analysed data and discussed the results and implications at all stages.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.K. (meike.kohler@icp.cat).

LETTER

doi:10.1038/nature11153

Widespread adoption of Bt cotton and insecticide decrease promotes biocontrol services

Yanhui Lu[1], Kongming Wu[1], Yuying Jiang[2], Yuyuan Guo[1] & Nicolas Desneux[3]

Over the past 16 years, vast plantings of transgenic crops producing insecticidal proteins from the bacterium *Bacillus thuringiensis* (Bt) have helped to control several major insect pests[1–5] and reduce the need for insecticide sprays[1,5,6]. Because broad-spectrum insecticides kill arthropod natural enemies that provide biological control of pests, the decrease in use of insecticide sprays associated with Bt crops could enhance biocontrol services[7–12]. However, this hypothesis has not been tested in terms of long-term landscape-level impacts[10]. On the basis of data from 1990 to 2010 at 36 sites in six provinces of northern China, we show here a marked increase in abundance of three types of generalist arthropod predators (ladybirds, lacewings and spiders) and a decreased abundance of aphid pests associated with widespread adoption of Bt cotton and reduced insecticide sprays in this crop. We also found evidence that the predators might provide additional biocontrol services spilling over from Bt cotton fields onto neighbouring crops (maize, peanut and soybean). Our work extends results from general studies evaluating ecological effects of Bt crops[1–4,6,12,13] by demonstrating that such crops can promote biocontrol services in agricultural landscapes.

Biological control is a valuable ecosystem service[14,15], but increasingly intensive farming strongly influences the populations of natural enemies and the biocontrol services they provide[16–18]. However, landscape biodiversity management and restricted use of pesticides may enhance biocontrol services in agro-ecosystems and could thus favour the development of sustainable farming[7–9]. Genetically engineered crops that express δ-endotoxins (Cry proteins) from *Bacillus thuringiensis* (Bt) have been increasingly implemented by farmers in many countries since 1996, and more than 6.6×10^7 ha of Bt crops were planted worldwide in 2011 (ref. 19). Bt crops have successfully controlled several major insect pests[1,2,4,5] and led to a drastic decrease in insecticide use on these crops[1,5,6]. Because insecticide applications have been gradually reduced in Bt crops, their widespread adoption may benefit natural enemies and may therefore potentially enhance associated ecosystem services such as the control of arthropod pests[10–12]. This last point has not yet been documented, especially with regard to the long-term landscape-level impacts[10].

From the 1970s, insecticides were applied extensively to control cotton bollworm (CBW), *Helicoverpa armigera*, the most serious insect pest on conventional cotton in China. However, control became almost impossible in the early 1990s because the pest became resistant to most insecticides, and unprecedented outbreaks in 1992 led to a wide overuse of insecticides. Consequently, in 1993, the Chinese government requested systematic insecticide applications in wheat crops for the control of the first-generation CBW; that is, before the following generations colonized cotton crops[20]. Although insecticide use decreased in cotton, this measure was not sustainable because insecticide applications were increased on wheat crops, resulting in both higher costs and environmental pollution. Bt cotton was therefore approved in 1997 for commercial use to control CBW, and it became the Chinese government's key measure against this cotton pest. It was rapidly planted on a large scale, rising to 2.4×10^6 ha by 2011 (more than 95% of the cotton crop in northern China). It managed CBW effectively, which led to decreased insecticide use on this pest[3,21].

The widespread adoption of Bt cotton may have favoured an increase in generalist natural enemy populations and promoted their associated biocontrol services. We therefore performed two assessments: first, whether implementing Bt cotton on a large scale induced an increase in populations of three groups of key generalist predators in China (ladybirds, lacewings and spiders) in both Bt cotton and three common neighbouring crops, namely maize, peanut and soybean; and second, whether this trend resulted in increased biocontrol services in agricultural landscapes in China. Aphids were selected as a pest model because they are common prey for generalist predators. During 1990–2011, research was conducted in six major cotton-growing provinces (Henan, Hebei, Shandong, Shanxi, Anhui and Jiangsu) in northern China, where about 2.6×10^6 ha of cotton and 3.3×10^7 ha of other crops (notably maize, peanut and soybean) are cultivated annually by more than ten million small-scale farmers.

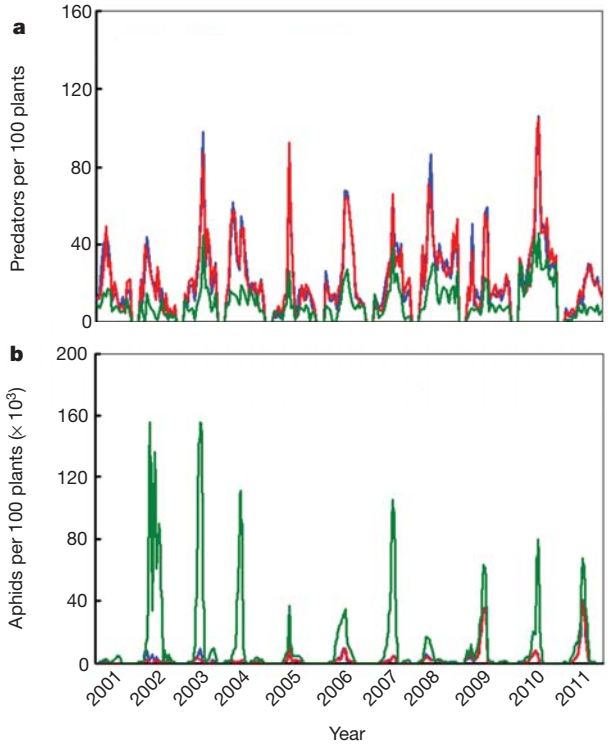


Figure 1 | Population densities of predators and aphids on cotton with different management regimes at Langfang experimental station (2001–2011). a, Predators. **b**, Aphids. The blue and red lines indicate Bt cotton and non-Bt cotton without insecticide sprays, respectively; the green line represents non-Bt cotton with CBW insecticide sprays (chemical control).

[1]State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Sciences, Beijing, 100193 China. [2]National Agro-Technical Extension and Service Center, Beijing, 100026 China. [3]French National Institute for Agricultural Research (INRA), UMR1355-ISA, 400 route des chappes, 06903 Sophia-Antipolis, France.

Predators and cotton aphids were sampled from 2001 to 2011 in Bt and non-Bt cotton plots at Langfang experimental station in Hebei province. No significant differences were found for predator ($P = 0.341$) and aphid ($P = 0.555$) abundances between Bt cotton and non-Bt cotton with similar management methods; that is, without application of insecticide (Fig. 1a, b and Supplementary Table 1a, b). However, predator abundance was significantly lower and aphid abundance was significantly higher in plots treated with insecticides for CBW management in comparison with insecticide-free plots ($P < 0.001$) (Fig. 1a, b and Supplementary Table 1a, b), although it varied over years (significant interactions between insecticide application and year). Bt cotton does not itself affect predator and aphid population levels[10,22], and generalist predators are clearly susceptible to broad-spectrum insecticides (such as synthetic pyrethroids) used against CBW. Thereafter, insecticide-induced aphid resurgence usually occurs with widespread applications of insecticides.

Predator abundance and insecticide use in cotton were monitored in 36 locations throughout northern China during 1990–2010 (Fig. 2a and Supplementary Table 2). Predator population levels gradually increased over that period, and relatively high population levels were always observed after Bt cotton was implemented in 1997 (Fig. 2b). In 14 selected locations, all three major groups of predators (ladybirds, lacewings and spiders) showed an increasing trend similar to that of the whole predator complex (Fig. 2b). Insecticide use patterns also changed greatly with Bt cotton implementation. After the introduction of Bt cotton, the number of insecticide sprays against CBW (and other insect pests in general), mainly pyrethroid and organophosphate insecticides (Supplementary Table 3), which have multiple negative effects on natural enemies[17], was lower than during the pre-Bt cotton period, namely 1990–1996 (Fig. 2c). Moreover, predator population level and number of insecticide sprays were positively and negatively related to Bt cotton planting proportions, respectively ($P < 0.001$; Supplementary Fig. 1a, b), and indicated the effect of its large-scale adoption on the predator population trend. Regression analyses showed that fewer insecticide sprays against CBW and all insect pests were correlated to a great extent with an increase in predator populations in northern China ($P < 0.001$) (Fig. 2d, e). The results were consistent in the six provinces, and insecticide use against CBW was a driving factor for predator population level in the cotton agroecosystem (all $P < 0.05$; Supplementary Table 4).

Cotton aphid abundance was surveyed in 24 locations from 1990 to 2010 (Supplementary Table 2) to assess the biocontrol services provided by generalist predators. Linear regression analyses showed that increasing generalist predator populations were correlated with decreasing aphid abundance in northern China in general ($P < 0.001$; Fig. 3a) and in all provinces except Shanxi (Supplementary Fig. 2a–e). During the three main periods studied—that is, without Bt cotton, with less than 90% and more than 90% of Bt cotton planting in the landscapes—aphid populations decreased significantly ($P < 0.001$; Fig. 3b). In addition, aphid population was negatively related to the proportion of Bt cotton planted ($P = 0.003$; Supplementary Fig. 3). Exclusion cage trials in 2010 and 2011 at Langfang and Xinxiang experimental stations (in

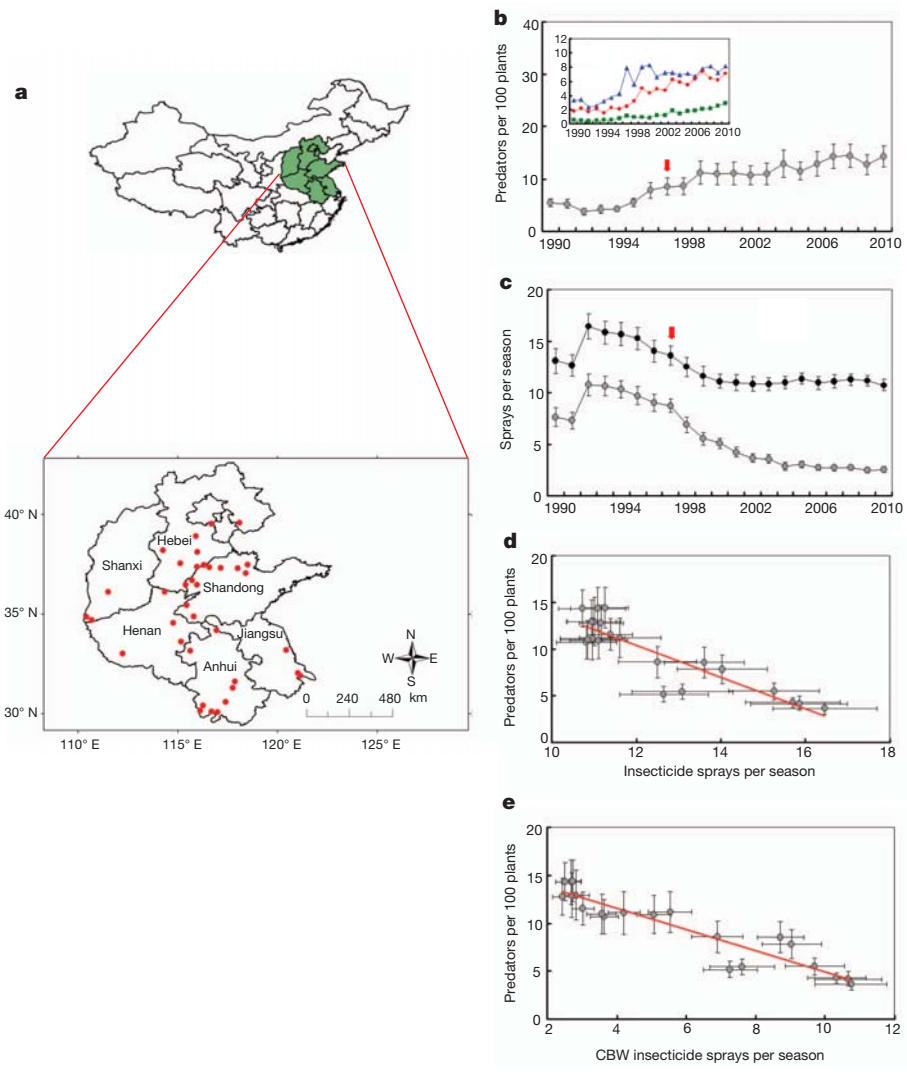


Figure 2 | Relationships between predator population density and number of insecticide sprays on cotton in northern China (1990–2010). **a**, Survey locations, indicated by red dots. **b**, Predator population density on cotton in commercial fields in 36 locations (each point represents one-year data; the red arrow indicates the beginning of Bt cotton use). Inset: population abundance of ladybirds (blue), spiders (red) and lacewings (green), collected from 14 locations. **c**, Number of insecticide sprays for CBW (grey points) and all insect pests (black points) on cotton; each point represents one-year data. **d**, Linear relationship between total number of insecticide applications, determined by pooling all treatments against all the insect pests on cotton (x), and the predator abundance (y) in cotton ($y = -1.69x + 30.63$, $F_{1,19} = 71.19$, $R^2 = 0.79$, $P < 0.0001$). **e**, Linear relationship between number of insecticide applications for CBW only (x) and predator abundance (y) ($y = -1.11x + 16.03$, $F_{1,19} = 137.32$, $R^2 = 0.88$, $P < 0.0001$). The data in **d** and **e** are replotted from **b** and **c**. All error bars show s.e.m.

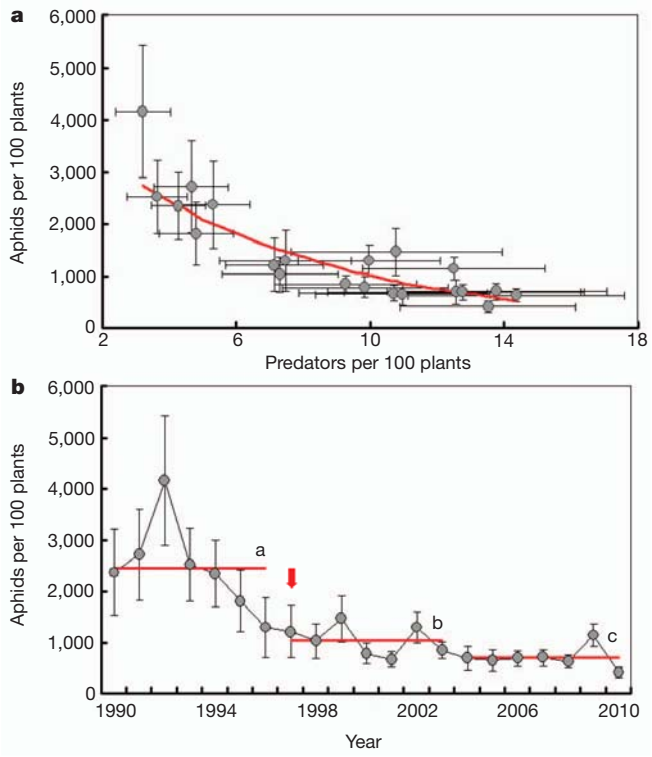


Figure 3 | Population abundance of cotton aphid in northern China (1990–2010) and relationship with predator abundance on cotton. **a**, Regression analysis between abundance of aphids (y) ($\log_e$-transformed) and predator abundance (x) ($y = e^{-0.15x + 8.39}$, $F_{1,19} = 69.67$, $R^2 = 0.79$, $P < 0.0001$). **b**, Aphid population density on cotton in commercial fields in 24 locations (each point represents one-year data, and the red arrow indicates the beginning of Bt cotton use). Red lines show the mean population density of aphids in cotton fields during three main periods, namely before Bt cotton planting (1990–1996), when Bt cotton planting was less than 90% of cotton surfaces planted (1997–2003) and when it was more than 90% (2004–2010). Red lines bearing different letters are significantly different at the $P < 0.05$ level in least-significant-difference post-hoc tests (one-way analysis of variance on $\log_e$-transformed data: $F_{2,18} = 27.57$, $P < 0.0001$). All error bars show s.e.m.

Hebei and Henan provinces, respectively) further demonstrated the significant effects of predators on aphid population growth in cotton fields (Supplementary Fig. 4a, b). As the cotton aphid populations declined, an invasive whitefly in cotton, *Bemisia tabaci*[20], probably served as an alternative prey for the increasing predator populations.

All these results indicate that the widespread adoption of Bt cotton ultimately promotes biocontrol services in the agroecosystem because decreased insecticide use leads to an increase in predator populations. Broadly speaking, measures that preserve predators in cotton fields greatly help to control aphid populations; for example, when insecticide applications in wheat were requested by the Chinese government (1993–1996) to prevent CBW outbreaks in cotton (see above), it led to a decreasing trend in aphid abundance (Fig. 3b).

Predator abundance was also monitored from 2001 to 2011 in three neighbouring crops: maize, peanut and soybean at Langfang experimental station. There was a positive relationship between predator abundance in cotton and soybean ($P = 0.019$; Fig. 4a), as well as between cotton and peanut (marginally significant, $P = 0.075$; Fig. 4b). We observed a similar trend in maize but it was not significant ($P = 0.216$; Fig. 4c). The increased predator abundance in maize was linked to a decrease in aphid pest abundance in that particular crop (marginally significant, $P = 0.061$; Fig. 4d).

Biocontrol services are important components in agro-ecosystems and could lead to the development of sustainable agriculture[7,15,23]. In conventional agricultural practices, insecticides are frequently used to control targeted pests, but they can lead to outbreaks of secondary pests by suppressing their natural enemies[24]. This so-called insecticide-induced resurgence was first reported for cotton aphid in the 1970s and was regarded as a key factor leading to population outbreaks of this pest in China[25]. Our work demonstrates the importance of natural enemies in the long-term suppression of the cotton aphid. The widespread adoption of Bt cotton, as a sustainable measure to reduce insecticide use, has indirectly promoted generalist predator abundance in Bt cotton fields but also to a smaller extent in three common adjacent crops in northern China. Bt crops therefore might enhance biocontrol services in agricultural landscapes through an increased abundance of generalist natural enemies. This study provides key information on long-term landscape-level ecological effects of Bt crops as well as useful insights, for example into the management of pest resurgence problems reported for many pests worldwide[26].

Generalist predators usually have great dispersal ability and can rely on various food sources. Hence, not only can they synchronously attack different insect pests in one field, but they can also colonize different habitats in different seasons[27,28]. Furthermore, some habitat management measures, such as inter-planting different crops or wild plants, have been adopted to provide resources such as food supply or shelter for natural enemies, thus increasing conservation biological control in adjacent fields[7,9,27,28]. We have demonstrated that decreasing insecticide application, through widespread Bt cotton plantings, sustained generalist predators and helped to suppress aphid populations in this

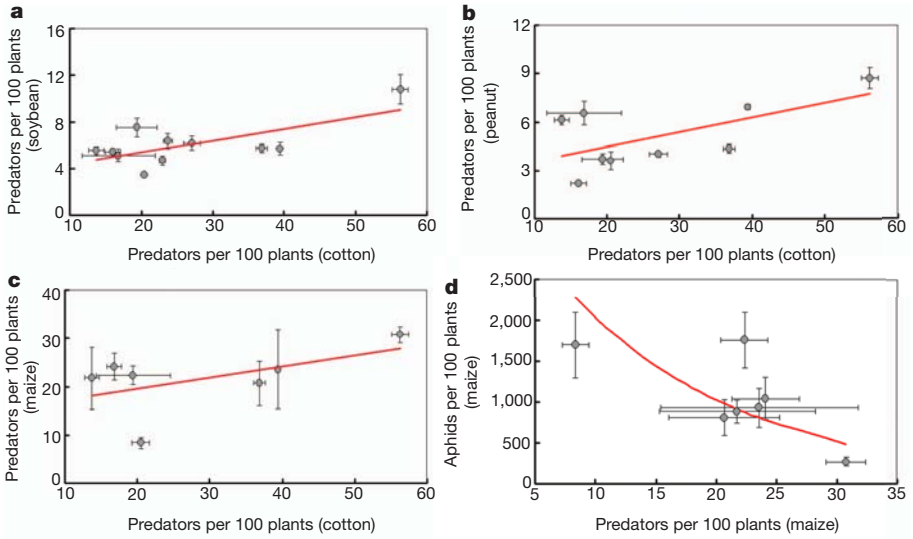


Figure 4 | Relationships between predator abundance on cotton and in three other crops, and between predator and aphid abundances in maize. Data on soybean (2001–2011), peanut (2001–2005 and 2008–2011) and maize (2001–2003 and 2008–2011) were collected at Langfang experimental station. **a**, Linear relationship between predator abundance on cotton (x) and on soybean (y) ($y = 0.10x + 3.38$, $F_{1,9} = 8.11$, $R^2 = 0.47$, $P = 0.0191$). **b**, Linear relationship between predator abundance on cotton (x) and on peanut (y) ($y = 0.09x + 2.66$, $F_{1,7} = 4.38$, $R^2 = 0.38$, $P = 0.0747$). **c**, Linear relationship between predator abundance on cotton (x) and on maize (y) ($y = 0.23x + 14.96$, $F_{1,5} = 2.00$, $R^2 = 0.29$, $P = 0.2164$). **d**, Relationship between predator abundance (x) and abundance of aphids in maize (y; $\log_e$-transformed data) ($y = e^{-0.07x + 8.31}$, $F_{1,5} = 5.80$, $R^2 = 0.54$, $P = 0.0610$). All error bars show s.e.m.

crop. Large-scale insecticide reduction is the key driver in such processes (for example see the period 1993–1996, during which insecticide decrease favoured an increase in predator populations and a decline of aphid populations). Higher generalist predator population levels in Bt cotton lead to lower insect pest levels in the crop, and these predators might provide additional biocontrol services spilling over from cotton fields onto neighbouring crops, although further work should be performed to document this last point. Broadly speaking, the deployment of Bt crops may favour biocontrol services and enhance economic benefits not only in Bt crop fields but also in the whole agricultural landscape. Field studies indicated that Bt crops protected natural enemies in comparison with non-Bt crops, which rely on conventional insecticides[22,29]. Our present study, demonstrating that biocontrol services are potentially provided by Bt crops throughout the agricultural landscape, may offer new options in developing conservation biological control measures at the landscape level.

Critical concerns about the ecological risk assessment of transgenic crops still remain, especially on a large scale[29]. The present study confirms no negative effects of one Bt crop, Bt cotton, on generalist predators in agricultural landscapes in China. More particularly, we have demonstrated a marked increase in generalist predator population levels and associated biocontrol services linked to decreased insecticide use owing to the widespread adoption of the Bt crop. Our work provides a comprehensive, long-term and large-scale assessment of the possible ecological and agricultural effects of transgenic crops.

METHODS SUMMARY

The study was based on large-scale surveys of predator and cotton aphid populations in cotton fields of northern China from 1990 to 2010 and on experiments and surveys that were performed at Langfang experimental station of the Chinese Academy of Agricultural Science (CAAS) during the period 2001–2011. The surveys and experiments focused on three major generalist predator groups (ladybirds, lacewings and spiders) and on aphid pests in cotton and in three common cotton-neighbouring crops, namely maize, peanut and soybean.

At the CAAS, we first assessed how cultural practices could affect predator and aphid populations in the long term in cotton fields; cotton plots were established every year and the abundance of predators and cotton aphids was surveyed in three different plot types: Bt cotton, non-Bt cotton and non-Bt cotton with insecticide. Second, we determined the impact of predators on aphid population in cotton by means of exclusion cage trials. Third, we evaluated the impact of implementing Bt cotton on predator and aphid populations in the neighbouring crops. Field plots were established in cotton, maize, peanut and soybean, and population dynamics of predators and aphids were monitored.

Large-scale surveys were conducted in six provinces in northern China (36 locations, 10–20 fields per location) to evaluate the impact of insecticide applications on the abundance of predators and aphids in cotton fields. We tested, first, the relationship between predator abundance and insecticide use during the period 1990–2010 (that is, including the period before and during the widespread adoption of Bt cotton by farmers), and second, how cotton aphid density was related to predator abundance during the same period.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 3 January; accepted 23 April 2012.
Published online 13 June 2012.

1. Shelton, A. M., Zhao, J. Z. & Roush, R. T. Economic, ecological, food safety, and social consequences of the deployment of Bt transgenic plants. *Annu. Rev. Entomol.* **47,** 845–881 (2002).
2. Carriere, Y. *et al.* Long-term regional suppression of pink bollworm by *Bacillus thuringiensis* cotton. *Proc. Natl Acad. Sci. USA* **100,** 1519–1523 (2006).
3. Wu, K. M., Lu, Y. H., Feng, H. Q., Jiang, Y. Y. & Zhao, J. Z. Suppression of cotton bollworm in multiple crops in China in areas with Bt toxin-containing cotton. *Science* **321,** 1676–1678 (2008).
4. Hutchison, W. D. *et al.* Areawide suppression of European corn borer with Bt maize reaps savings to non-Bt maize growers. *Science* **330,** 222–225 (2010).
5. Tabashnik, B. E. *et al.* Suppressing resistance to Bt cotton with sterile insect releases. *Nature Biotechnol.* **28,** 1304–1307 (2010).
6. Cattaneo, M. G. *et al.* Farm-scale evaluation of the impacts of transgenic cotton on biodiversity, pesticide use, and yield. *Proc. Natl Acad. Sci. USA* **103,** 7571–7576 (2006).
7. Landis, D. A., Wratten, S. D. & Gurr, G. M. Habitat management to conserve natural enemies of arthropod pests in agriculture. *Annu. Rev. Entomol.* **45,** 175–201 (2000).
8. Crowder, D. W., Northfield, T. D., Strand, M. R. & Snyder, W. E. Organic agriculture promotes evenness and natural pest control. *Nature* **466,** 109–112 (2010).
9. Winqvist, C. *et al.* Mixed effects of organic farming and landscape complexity on farmland biodiversity and biological control potential across Europe. *J. Appl. Ecol.* **48,** 570–579 (2011).
10. Romeis, J. *et al.* Transgenic crops expressing *Bacillus thuringiensis* toxins and biological control. *Nature Biotechnol.* **24,** 63–71 (2006).
11. Bale, J. S., van Lenteren, J. C. & Bigler, F. Biological control and sustainable food production. *Phil. Trans. R. Soc. B* **363,** 761–776 (2008).
12. Gatehouse, A. M. R., Ferry, N., Edwards, M. G. & Bell, H. A. Insect-resistant biotech crops and their impacts on beneficial arthropods. *Phil. Trans. R. Soc. B* **366,** 1438–1452 (2011).
13. Wolfenbarger, L. L., Naranjo, S. E., Lundgren, J. G., Bitzer, R. J. & Watrud, L. S. Bt crop effects on functional guilds of non-target arthropods: a meta-analysis. *PLoS ONE* 3**,** e2118 10.1371/journal.pone.0002118 (2008).
14. Bianchi, F. J. J. A., Booij, C. J. H. & Tscharntke, T. Sustainable pest regulation in agricultural landscapes: a review on landscape composition, biodiversity, and natural pest control. *Proc. R. Soc. Lond. B* **273,** 1715–1727 (2006).
15. Losey, J. E. & Vaughan, M. The economic value of ecological services provided by insects. *Bioscience* **56,** 311–323 (2006).
16. Chapin, F. S. *et al.* Consequences of changing biodiversity. *Nature* **405,** 234–242 (2000).
17. Desneux, N., Decourtye, A. & Delpuech, J. M. The sublethal effecs of pesticides on beneficial arthropods. *Annu. Rev. Entomol.* **52,** 81–106 (2007).
18. Landis, D. A., Gardiner, M. M., van der Werf, W. & Swinton, S. M. Increasing corn for biofuel production reduces biocontrol services in agricultural landscapes. *Proc. Natl Acad. Sci. USA* **105,** 20552–20557 (2008).
19. James, C. *Global Status of Commercialized Biotech/GM crops: 2011* (ISAAA brief no. 43, International Service for the Acquisition of Agri-Biotech Applications, 2011).
20. Wu, K. M. & Guo, Y. Y. The evolution of cotton pest management practices in China. *Annu. Rev. Entomol.* **50,** 31–52 (2005).
21. Lu, Y. H. *et al.* Mirid bug outbreaks in multiple crops correlated with wide-scale adoption of Bt cotton in China. *Science* **328,** 1151–1154 (2010).
22. Naranjo, S. E. Long-term assessment of the effects of transgenic Bt cotton on the abundance of nontarget arthropod natural enemies. *Environ. Entomol.* **34,** 1193–1210 (2005).
23. Ragsdale, D. W., Landis, D. A., Brodeur, J., Heimpel, G. E. & Desneux, N. Ecology and management of the soybean aphid in North America. *Annu. Rev. Entomol.* **56,** 375–399 (2011).
24. Hardin, M. R. *et al.* Arthropod pest resurgence: an overview of potential mechanisms. *Crop Prot.* **14,** 3–18 (1995).
25. Wu, K. W. & Liu, Q. X. Study on the resurgence caused by insecticides for cotton aphid, *Aphis gossypii. Acta Ecol. Sin.* **12,** 341–347 (1992).
26. Rao, C. N., Shivankar, V. J. & Singh, S. in *Encyclopedia of Pest Management* Vol 2 (ed. Pimentel, D.) 597–601 (CRC Press, 2007).
27. Symondson, W. O. C., Sunderland, K. D. & Greenstone, H. M. Can generalist predators be effective biocontrol agents? *Annu. Rev. Entomol.* **47,** 561–594 (2002).
28. Tscharntke, T. *et al.* Conservation biological control and enemy diversity on a landscape scale. *Biol. Control* **43,** 294–309 (2007).
29. Romeis, J., Shelton, A. M. & Kennedy, G. G. *Integration of Insect-resistant Genetically Modified Crops within IPM Programs* (Springer, 2008).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the agricultural technicians of the 36 surveyed locations for providing data sets from cotton fields; W. Li and L. Wang for providing the predator exclusion cage data sets; Y. Zhang and C. Li for assistance in making the survey map; H. Yuan for help in preparing the list of insecticides in cotton; and R. Senoussi for advice on data analyses. This work was funded by the Key Project for Breeding Genetically Modified Organisms (grant no. 2011ZX08012-004) and the International Science and Technology Cooperation Project (grant no. 2010DFA32200).

Author Contributions K.W., Y.L. and Y.G. designed and performed the experiments. Y.J. performed the surveys. Y.L., K.W. and N.D. analysed the data and shared in the scoping and writing responsibilities.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to K.W. (kmwu@ippcaas.cn).

METHODS

Aphid pests and predator complex considered in the study. In northern China, several aphid species are reported as pests on cotton, maize, peanut and soybean. *Aphis gossypii* Glover (cotton aphid) is the main aphid pest in cotton fields in northern China where there are two key biotypes (called seedling and summer aphids, respectively). Only the summer aphid, which colonizes fields from early July to late August, is considered a major cotton pest[20,30,31]. The seedling aphid is controlled by insecticide treatments applied on the seeds; these compounds do not last long enough within plants to provide control of the summer aphid[20]. *Rhopalosiphum maidis* Fitch, *R. padi* Linnaeus and *Sitobion avenae* Fabricius are the three dominant aphid pest species on maize[31], and *A. glycines* Matsumura and *A. craccivora* Koch are the main aphid species on soybean and peanut, respectively[31]. In our study, the above aphid species were considered in assessing the biocontrol services provided by generalist predators.

In northern China, there are three dominant groups of generalist predators in cotton field (more than 90% of all the predators[30]): ladybirds, lacewings and spiders. In our study we therefore focused on a predator complex composed of ladybirds (*Propylea japonica* Thunberg, *Harmonia axyridis* Pallas, *Coccinella septempunctata* L. and *Adonia variegata* Goeze), lacewings (*Chrysopa septempunctata* Wesmael, *Chrysoperla sinica* Tjeder and *Chrysopa formosa* Brauer) and spiders (*Erigonidium graminicolum* Sundevall, *Misumenopos tricuspidata* Fahricius and *Pardosa t-insignita* Boes. et Str.); these compose the most common predators in agricultural landscape of that region. These groups of predators are also common in maize, peanut and soybean fields and were thus also considered as a predator complex for these three crops[32–34].

Impact of agricultural practices on predator and cotton aphid populations. Survey experiments were conducted from 2001 to 2011 at Langfang experimental station (39.53° N, 116.70° E), Chinese Academy of Agricultural Sciences (CAAS), Hebei province, China. Fifteen cotton plots (400 m^2 each) were established every year and were managed with agronomic practices that are standard in northern China. A randomized block design with three replicates was used, which included two Bt and two conventional cotton varieties. One Bt cotton variety expressing *Cry1Ac* (NuCOTN33B) and another Bt cotton varieties expressing *Cry1A* (SGK321) were supplied by Monsanto Co. and the Biotechnology Research Institute, CAAS, respectively. Two conventional cotton varieties (Shiyuan321 and Zhong12) were obtained from the Institute of Plant Protection, CAAS. Shiyuan321 was the non-transgenic isoline of SGK321. Every year, the trial consisted of three treatments: Bt cotton and non-Bt cotton plots without insecticide, and non-Bt cotton (one variety, Zhong12) plots with insecticides. β-Cypermethrin (pyrethroid) and phoxim (organophosphate) were used when insecticides were applied. The choice of these two insecticides, and their frequency of application, were both based on management guidelines for CBW (*Helicoverpa armigera*) used throughout the early 1990s in northern China[20] (Supplementary Table 3).

The abundances of predators and cotton aphid were surveyed in the three cotton plot types (Bt cotton, non-Bt cotton and non-Bt cotton with insecticide) every 4 or 5 days from mid-June to late August from 2001 to 2011. At each sampling date, 100 plants at five random locations per plot[35] were visually inspected and all predators and aphids were recorded. No significant differences ($P > 0.05$) were found between cotton varieties, so we combined NuCOTN33B and SGK321 as Bt cotton, and Shiyuan321 and Zhong12 as non-Bt cotton, for further analysis. A three-way ANOVA was used to analyse the effects of the cotton variety (Bt cotton and non-Bt cotton), insecticide treatments (chemical control and non-chemical control) and sampling year on predator and aphid abundance, and the interactions between year and cotton variety, and between year and insecticide spray; the means were compared by the least-significant-difference (LSD) test at $P = 0.05$.

Survey of predators and cotton aphid in cotton crops in northern China. From 1990 until 2010, commercial cotton fields in 36 locations in six provinces (Henan, Hebei, Shandong, Shanxi, Anhui and Jiangsu) of northern China were surveyed for predators and cotton aphid (Supplementary Table 2). Insect populations were recorded every 3–10 days from early June to late August every year. For each survey, 10–20 cotton fields were sampled per location. Within each field, a total of 50–100 cotton plants at five random locations were visually inspected for predators[30]. Among the 36 locations, ladybirds, lacewings and spiders were recorded as a predator complex in 22 locations, whereas in 14 sites the three predator types were recorded individually. The 14 locations included Anxin and Xinji from Hebei province; Dezhou, Binzhou and Chengwu from Shandong province; Ruicheng, Yongji and Linfen from Shanxi province; Dongzhi, Wangjiang and Taihu from Anhui province; and Dafeng, Tongzhou and Haimen from Jiangsu province. The cotton aphid populations were surveyed in 24 locations in five provinces (Supplementary Table 2), using the same sampling schedule as for survey of predators[30]. On each plant, an upper leaf, a middle leaf and a lower leaf were examined for aphid presence. At the same time, all insecticide applications (for management of CBW and other arthropod pests) were recorded per field per year.

Linear regression analyses were used to assess the relationship between predator abundance and insecticide use on the data set gathered from 1990 to 2010—that is, including the period during which Bt cotton was increasingly adopted in China by farmers. Both simple and forward stepwise regressions were used to relate predator abundance and insecticide use against CBW and all insect pests for each province and the whole of northern China in the 1990–2010 data set. Simple linear models were used to assess the relationship between aphid density (log-transformed) and predator abundance from early July to late August (in the 1990–2010 data set) for each province and for the whole of northern China. Linear regression analyses were used to assess the relationship between predator abundance and aphid abundance (log-transformed) with Bt cotton planting proportions. In this analysis, the mean abundances of predators and aphids during 1990–1996 were included as the data when the Bt cotton planting proportion was 0.

To evaluate the impact of the predators on cotton aphid population further, exclusion cage trials were conducted in 2010 and 2011 at Langfang experimental station and Xinxiang experimental station of CAAS (Henan Province, 35.09° N, 113.48° E). This trial included a caged treatment and an open-field treatment as control[18,36]. The cage was 2 m wide by 2 m wide by 1.5 m high and made from the insect mesh net, which allowed the emigration and immigration of alate aphids and its parasitoids, but blocked the predators[37]. Ten cotton plants were covered in each cage. This trial began in July, when almost only apterous aphids were in the field[38], and was limited to 15 days to prevent the appearance of alate aphids in the cage[18,36]. At each site, cage treatments with three or four replicates were established when aphid density reached an average of 2 individuals per plant in 2010 and 20 individuals per plant in 2011. We recorded the aphid abundance 15 days after treatment. Meanwhile, predator densities were surveyed three times, on day 0, day 5 and day 10, in ten randomly selected cotton plants in open field during the whole trial. The aphid abundance in caged and open plants was compared by one-way ANOVA followed by a post-hoc LSD test. Before analysis, the data for aphid abundance were $\log_e$-transformed.

Impact of Bt cotton adoption on populations of predators in neighbouring crops. Population dynamics of the predator complex were monitored in cotton, soybean, peanut and maize field plots from 2001 to 2011 (except for maize, which was monitored during 2001–2003 and 2008–2011, and soybean, which was monitored during 2001–2006 and 2008–2011) at Langfang experimental station. Every year, a total of nine field plots (400 m^2 each) were established for each crop type and they were managed in the same way, applying the same fertilizers and irrigation treatment, free of any pesticide. A randomized block design with three replicates for each crop type was used. One Bt cotton variety, SGK321, was supplied by the Biotechnology Research Institute (CAAS); the maize (var. Shengshi29), soybean (var. Zhonghuang13) and peanut (var. Huayu16) were provided by Langfang experimental station (CAAS), the Institute of Crop Sciences (CAAS) and the Shandong Peanut Research Institute, respectively. The abundance of predators was recorded in the four different crops (Bt cotton, maize, peanut and soybean) every four or five days from mid-June to late August. At each sampling date, 100 plants in five random spots per plot were visually inspected and all predators were recorded. Linear regression analyses were used to assess the relationship between seasonal density of predators on cotton and soybean/peanut (data set covering the 2001–2011 period) and maize (data set covering the 2001–2003 and 2008–2011 periods).

For the maize plots, maize aphids were also recorded because they are well known as the main pests on maize in northern China. Population levels of aphids on soybean and peanut crops at Langfang experimental station were very low during the course of our study and therefore the data could not be considered in the framework of the study. A simple linear model was used to assess the relationship between aphid abundance (log-transformed) and predator abundance on maize.

30. Qu, X. F. *Cotton Pest Forecast in China: the Criterion, Zoning and Method* (China Scien-tech Press, 1992).
31. Institute of Plant Protection, Chinese Academy of Agricultural Sciences. *Crop Diseases and Pests in China* (China Agricultural Press, 1995).
32. Xu, H. F., Mu, S. M., Xu, Y. Y., Mu, J. Y. & Dong, C. X. On the community structure of major insect pests and natural enemies in summer corn field inlaid in cotton area. *Acta Phytophyl. Sin.* **27,** 199–204 (2000).
33. Yang, Q. M., Sun, M., Xu, Y. F., Shi, A. J. & Mu, J. Y. On the community structure of major insect pests and natural enemies in summer bean field. *J. Shandong Agric. Univ.* **35,** 217–220 (2004).
34. Ding, Y. Q. & Cheng, S. L. Preliminary investigation and utilization of natural enemies of aphid in peanut field. *J. Peanut Sci.* **39,** 45–47 (2010).

35. Wu, K. M. & Guo, Y. Y. Influences of *Bacillus thuringiensis* Berliner cotton planting on population dynamics of the cotton aphid, *Aphis gossypii* Glover, in northern China. *Environ. Entomol.* **32,** 312–318 (2003).
36. Gardiner, M. M. *et al.* Landscape diversity enhances biological control of an introduced crop pest in the north-central USA. *Ecol. Appl.* **19,** 143–154 (2009).
37. Miao, J., Wu, K. M., Hopper, K. R. & Li, G. X. Population dynamics of *Aphis glycines* (Homoptera: Aphididae) and impact of natural enemies in northern China. *Environ. Entomol.* **36,** 840–848 (2007).
38. Lu, Y. H., Qi, F. J. & Zhang, Y. J. *Integrated Management of Diseases and Insect Pests in Cotton* (Golden Shield Press, 2010).

LETTER

doi:10.1038/nature11144

Ecological opportunity and sexual selection together predict adaptive radiation

Catherine E. Wagner[1,2,3,4], Luke J. Harmon[5] & Ole Seehausen[1,2]

A fundamental challenge to our understanding of biodiversity is to explain why some groups of species undergo adaptive radiations, diversifying extensively into many and varied species, whereas others do not[1,2]. Both extrinsic environmental factors (for example, resource availability, climate) and intrinsic lineage-specific traits (for example, behavioural or morphological traits, genetic architecture) influence diversification, but few studies have addressed how such factors interact. Radiations of cichlid fishes in the African Great Lakes provide some of the most dramatic cases of species diversification. However, most cichlid lineages in African lakes have not undergone adaptive radiations. Here we compile data on cichlid colonization and diversification in 46 African lakes, along with lake environmental features and information about the traits of colonizing cichlid lineages, to investigate why adaptive radiation does and does not occur. We find that extrinsic environmental factors related to ecological opportunity and intrinsic lineage-specific traits related to sexual selection both strongly influence whether cichlids radiate. Cichlids are more likely to radiate in deep lakes, in regions with more incident solar radiation and in lakes where there has been more time for diversification. Weak or negative associations between diversification and lake surface area indicate that cichlid speciation is not constrained by area, in contrast to diversification in many terrestrial taxa[3]. Among the suite of intrinsic traits that we investigate, sexual dichromatism, a surrogate for the intensity of sexual selection, is consistently positively associated with diversification. Thus, for cichlids, it is the coincidence between ecological opportunity and sexual selection that best predicts whether adaptive radiation will occur. These findings suggest that adaptive radiation is predictable, but only when species traits and environmental factors are jointly considered.

Adaptive radiations are iconic systems for the study of evolutionary processes because they generate a wealth of ecological and species diversity, often on very rapid timescales[2,4]. Some of the most spectacular examples of young adaptive radiations occur on oceanic islands or in lakes, but such geographically circumscribed habitats are no guarantee for a radiation to evolve. Why is it that some lineages diversify markedly, whereas closely related lineages in the same habitat do not?

One point of view is that adaptive radiation is a consequence of newly arising ecological opportunity[1,4]. Extrinsic ecological factors that have been linked to adaptive radiation include a paucity of competing lineages[1,2], predation regime[5], biotic insularity[6], habitat complexity[7] and habitat area[3]. In addition, latitude[8] and energy (measured as solar radiation or primary productivity)[9] have been classically linked to variation in broad-scale patterns of diversity (for example, the latitudinal diversity gradient), but these factors have not been previously investigated in the context of adaptive radiations.

Another point of view is that differences in diversification result primarily from variation in lineage-specific traits that affect speciation rates, such as prevalence of sexual selection[10], ecological specialization[11], ecological versatility[12] and spatial vagility[3]. There is mounting evidence for traits underlying variation in diversification rates, but the overall proportion of variation explained is generally low[13]. A main challenge in explaining the causes of diversification lies in identifying the relative roles of intrinsic and extrinsic factors, and how these factors interact to determine the rate and volume of species radiations.

Rarely have the influences of multiple extrinsic and intrinsic factors been considered simultaneously in the study of adaptive radiation. Since the discovery of the species-rich African lake cichlid faunas, hypotheses for the spectacular diversity of these fishes have proliferated, invoking environmental factors[14,15], intrinsic traits[12,16] and their interactions[14,17] as influences on radiation. However, these hypotheses remain untested at macroevolutionary scales. Most research has focused on the cichlid radiations in Lakes Victoria, Malawi and Tanganyika, but cichlids have independently diversified within African lakes on more than 30 occasions, and have colonized lakes without diversifying on more than 120 occasions. These replicated cases of both occurrence and absence of diversification provide an opportunity to test which factors predict whether a cichlid lineage will diversify.

We built a molecular phylogeny for African cichlids (Supplementary Information 1), and placed all lacustrine African cichlids included in our data set on this tree (Fig. 1). We then collated information on lake characteristics for 46 lakes harbouring cichlids across the African continent, including lake depth, surface area, net solar radiation (hereafter 'energy'), latitude, elevation, the presence of predatory fishes and time for diversification (Fig. 1 and Supplementary Information 2). We collected data on intrinsic traits of cichlid lineages, including the presence of a polygamous mating system, mouthbrooding, generalized egg dummies and/or morphologically derived 'haplochromine' egg dummies[18] (used in courtship and in fertilization of eggs in the mouth of the female) and sexual dichromatism. We then tested for associations between these predictor variables and cichlid 'diversification state': that is, whether a lineage has diversified upon entering a lake or has failed to do so. We conducted analyses using two thresholds for the endemic diversity required to qualify as a radiation: at the lower threshold, we counted any lineage that had undergone at least one intralacustrine speciation event as radiating; at the higher threshold, we counted lineages as radiating only if they produced five or more endemic species within a single lake. Furthermore, we tested jointly for factors predicting radiation and the species richness of radiations using phylogenetic hurdle Poisson regression (see Supplementary Information 5.3).

We examined relationships between cichlid radiation and single predictor variables (Supplementary Tables 3 and 4) and then assessed the combined influence of predictor variables on diversification state in multiple regression models using Akaike information criterion-based model averaging corrected for small sample sizes (AICc)[19] followed by phylogenetic multiple logistic regression and phylogenetic hurdle Poisson regression of a reduced predictor variable set. The best-supported predictor variables in our multiple regression models

[1]Department of Fish Ecology & Evolution, EAWAG Centre for Ecology, Evolution and Biogeochemistry, 6047 Kastanienbaum, Switzerland. [2]Department of Aquatic Ecology, Institute of Ecology and Evolution, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland. [3]Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853, USA. [4]Fuller Evolutionary Biology Program, Cornell University Lab of Ornithology, Ithaca, New York 14850, USA. [5]Department of Biological Sciences, University of Idaho, Moscow, Idaho 83844, USA.

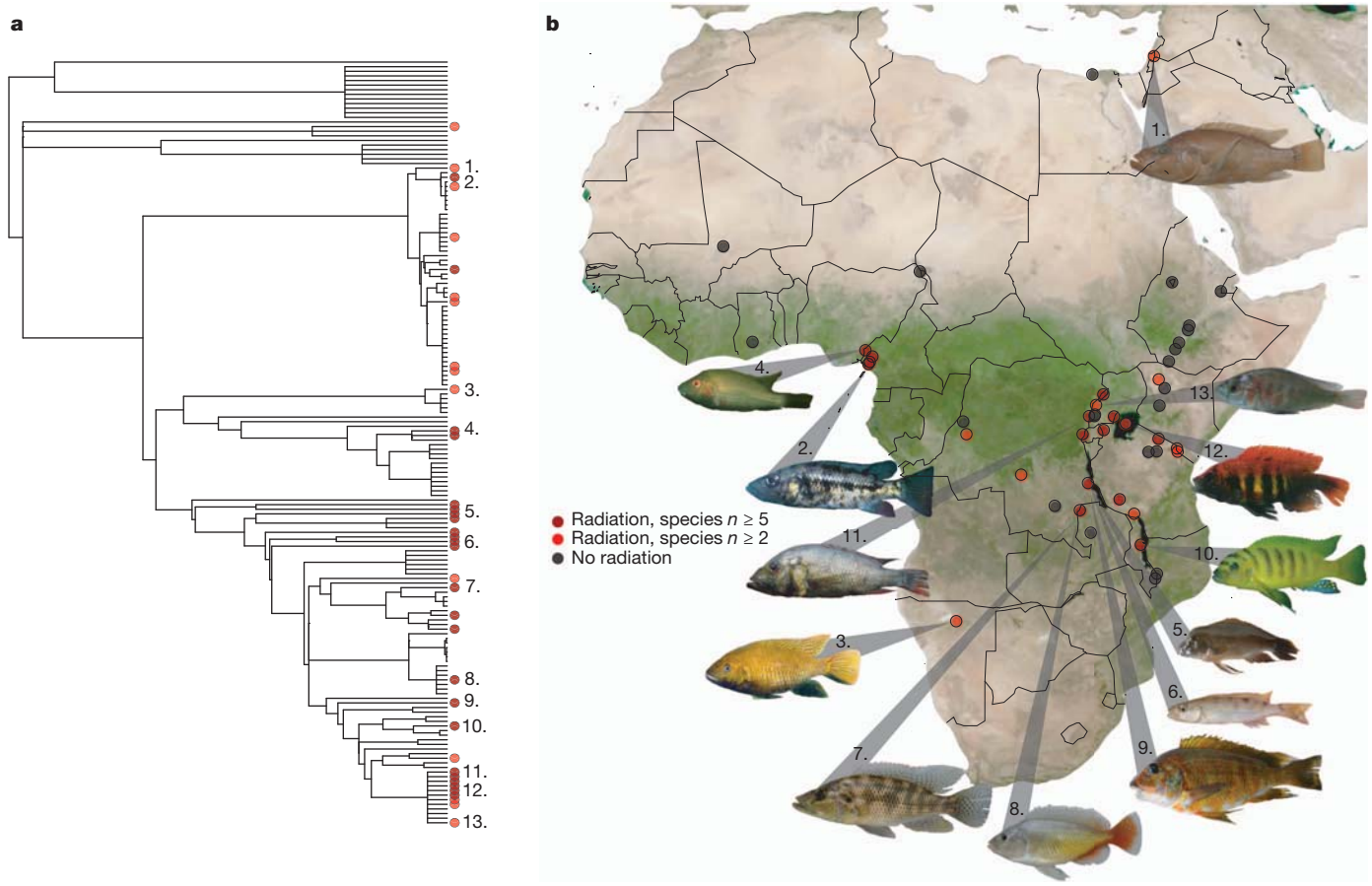


Figure 1 | Cichlid diversification is phylogenetically and geographically widespread. a, The distribution of intralacustrine adaptive radiation across the African cichlid phylogeny. Each tip represents one lineage in a lake; light red dots indicate at least one intralacustrine speciation event, dark red dots indicate radiation of five or more species. **b**, The geographic distribution of cichlid diversification in lakes across Africa. Each dot represents a lake; grey dots indicate colonization but no diversification within the lake; light red dots indicate the presence of at least one lineage with at least one intralacustrine speciation event; dark red dots indicate at least one lineage with five or more species formed within the lake.

include both environmental variables and lineage-specific traits (Fig. 2). Lake depth, energy and sexual dichromatism are the most consistently well-supported predictor variables: all were positively related to diversification with high relative importance scores for both of our radiation thresholds (that is, ≥ 2 or ≥ 5) in logistic regressions, and were top predictors in the binary portion of hurdle Poisson regressions (Supplementary Fig. 6). There was equivocal support for diversification to be more likely in lakes with small surface area, because this variable only had a high relative importance score at the lower threshold of radiation in logistic regression analyses. As a conservative test, we conducted these analyses excluding Lake Tanganyika, an outlier in both depth and age. In these tests, the same environmental variables were the strongest extrinsic factors associated with diversification, and a negative effect of elevation emerged as a predictor of diversification for both thresholds. Sexual dichromatism remained the most consistent intrinsic trait predictor of diversification (Fig. 2 and Supplementary Fig. 6).

All regression models show a strong association between lake depth and cichlid diversification, which is consistent with depth being an important axis of niche differentiation in intralacustrine speciation in fishes. Lake depth and age are typically highly correlated (Supplementary Information 3), and deeper lakes might additionally have greater environmental stability and/or greater persistence times, both of which would allow lineages more opportunity for diversification. Results showing a relationship between cichlid radiation and time for diversification support this idea (Supplementary Information 4). However, analyses on a subset of the data wherein time and depth were uncorrelated show a better fit for lake depth than time for diversification in predicting radiation (Supplementary Information 3), suggesting a role for depth apart from time in cichlid radiation. Depth partitioning of resources and reproduction is important in many cases of speciation in fishes[20], and case studies indicate that depth-specific divergence in mating traits and preferences and depth-specific ecological adaptation can be key factors in cichlid speciation[21]. Furthermore, increased depth increases habitat area for fishes, and the resultant larger population sizes may influence speciation and extinction rates.

Net solar radiation emerges as a second strong predictor of cichlid diversification in multiple regression models. Links between energy and evolutionary diversification have been frequently proposed in the context of latitudinal gradients in species richness[8,9], although only rarely has this relationship been tested. Increased energy input might increase carrying capacities, leading to larger total population sizes and increased rates of speciation and/or lower rates of extinction. Alternatively, high inputs of energy may lead to shortened generation times and/or increased mutation rates, resulting in increased rates of population differentiation and speciation[8,9].

In contrast to diversification in terrestrial systems[3], we find that increased lake surface area does not increase the likelihood that colonizing lineages will undergo intralacustrine speciation. Ascertainment bias could influence this result: data on the presence of species in very small lakes are rarer than in large lakes, and many of the small lakes included in our data set are known because they harbour endemic cichlids (Supplementary Information 5.2). However, in other systems such as *Anolis* lizards, very small islands never host adaptive radiations[22]. Regardless of potential size-related sampling bias, this finding demonstrates a marked contrast between cichlids and terrestrial taxa in that speciation is apparently not constrained by surface area in cichlids.

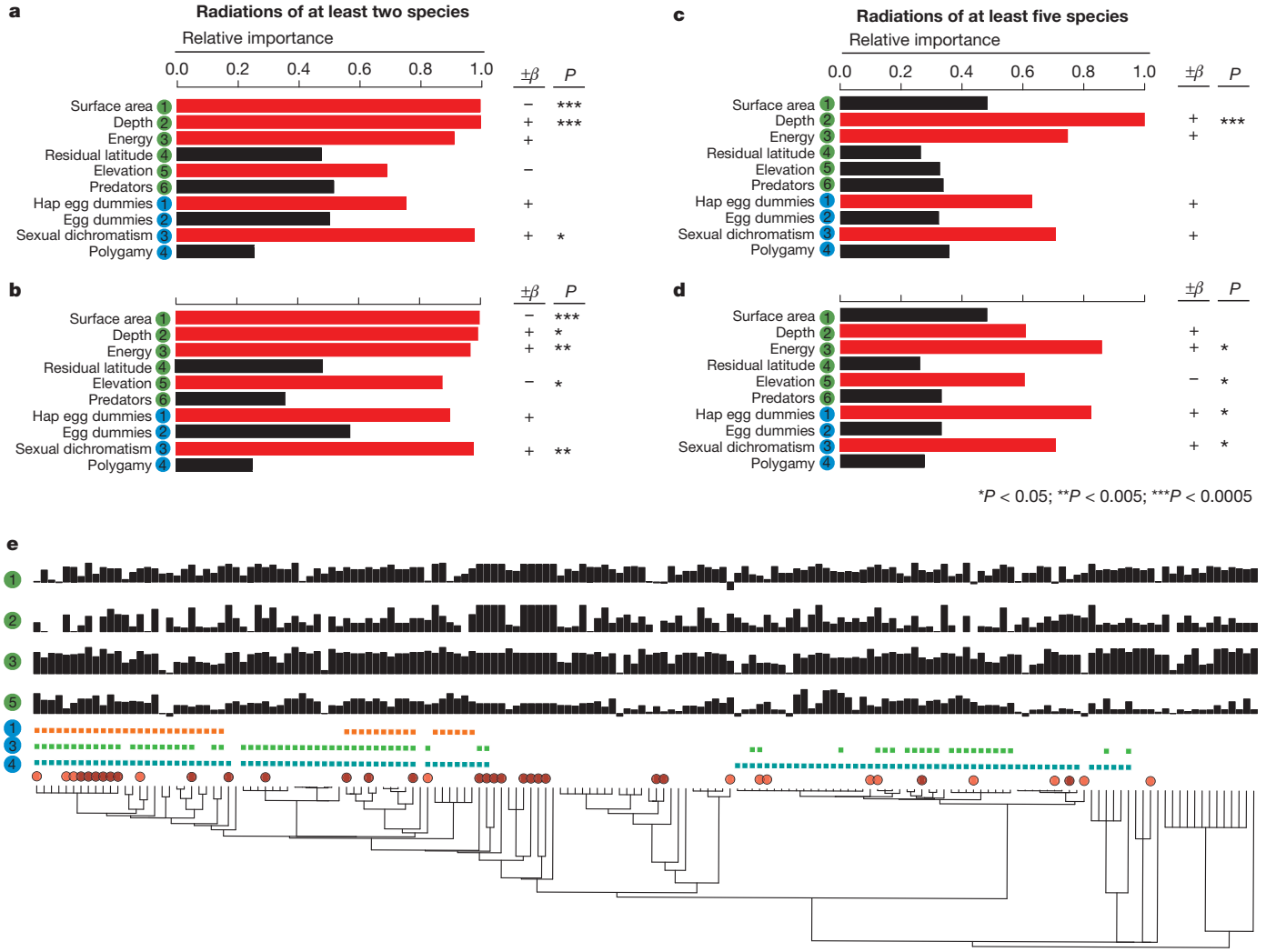


Figure 2 | Multiple logistic regression shows that environment and lineage-specific traits together best explain cichlid diversification in African lakes. **a**–**d**, Bar length is proportional to relative importance values, with red bars indicating relative importance values greater than 0.6 in multiple regression models. **a**, **c**, Results for the full data set. **b**, **d**, Results excluding Lake Tanganyika. Plus and minus symbols indicate the sign of multiple logistic regression coefficient estimates (β). Asterisks indicate the significance of a term in phylogenetic multiple logistic regression analyses. Green and blue circle labels represent environmental variables and species traits, respectively. Among environmental variables, there are positive associations between diversification and lake depth and environmental energy. Lake surface area is a negative predictor of diversification when radiations are considered to consist of two or more endemic species (**a**, **b**), but for larger radiations (five or more species) the significance of this negative size effect disappears (**c**, **d**); extremely species-rich radiations only occur in large lakes. Among lineage-specific traits, the presence of sexual dichromatism is a consistent predictor of diversification. Haplochromine egg dummies are consistently associated with diversification, but in most cases the significance of this effect disappears when phylogeny is accounted for. **e**, High relative importance variables, plus the low relative importance variable polygamous mating system, plotted on the African cichlid phylogeny (dots indicate radiation, as in Fig. 1). All cichlid lineages with sexual dichromatism or haplochromine egg dummies have polygamous mating systems. That the evolution of sexual dichromatism and egg dummies only occurs in lineages with polygamous mating systems suggests that polygamy is a prerequisite to strong sexual selection in cichlids.

The positive association between sexual dichromatism and diversification in all our models (Fig. 2 and Supplementary Information 4 and 5) suggests that the intensity of sexual selection may be a key influence on the probability that lineages radiate. Sexual dichromatism is a common proxy for strong sexual selection in studies of diversification[10]. Variation among and within populations in traits under sexual selection, and in associated preferences, can readily lead to pre-mating isolation among populations, and thereby facilitate speciation[23,24]. Sexual selection is known to be important in cichlid speciation from case studies, but here we show an association between sexual selection and diversification in cichlids at macroevolutionary scales. Examination of the co-occurrence between dichromatism and mating system shows that sexual dichromatism only evolves in lineages that have polygamous mating systems (Fig. 2), a pattern predicted if mating system determines opportunity for sexual selection[25]. Yet, mating system does not emerge as an important predictor of radiation in our models. This result suggests that dichromatism is a more direct indicator of the actual strength of sexual selection than is mating system, a pattern that has been suggested in meta-analysis of findings from other taxa[10] but which has never been tested in cichlids.

Although African cichlid fishes are an iconic example of adaptive radiation, our analysis shows great heterogeneity in the occurrence of adaptive radiation across this clade: most lineages present in lakes do not diversify. However, some lineage traits significantly predict whether radiations happen when a suitable environment is colonized. This result makes clear that the propensity for high diversification is not an intrinsic property of all cichlids, but one that has evolved in some branches of the cichlid tree. Although other unmeasured variables undoubtedly explain additional variation in the occurrence of adaptive radiation across the African cichlid phylogeny, we show here,

that cichlid adaptive radiations are not a simple function of any one predictor variable, but instead are best predicted by variables representing both extrinsic environmental effects and intrinsic, lineage-specific traits. For cichlids, it is the combined effects of the intensity of sexual selection and environmental opportunity, in the form of lake depth, energy availability and lake age, that best predict whether adaptive radiation will occur. More generally, the finding that propensity for adaptive radiation is underlain by several factors helps to explain why only some taxa radiate, even in environmental settings—such as islands and lakes—that are home to some of evolution's classic cases of adaptive radiation. Thus it is possible that adaptive radiation is predictable, but only when traits and environmental factors are jointly considered.

METHODS SUMMARY

We built maximum likelihood phylogenies in RAxML[26] using nine genes and sequences from 656 African cichlid species, and used PATHd8 (ref. 27), three geological dates and one fossil date to time-calibrate these trees (see Supplementary Information 1).

We compiled information about the presence and species richness of cichlid lineages in lakes across Africa. Because most colonizing lineages do not diversify, this data set is zero-inflated, and thus for analysis in a logistic regression framework we coded lineages in each lake as either 'non-diversifying' or 'diversifying' using one of two thresholds (Supplementary Information 2.3–4). We then compiled information about character states for traits potentially linked to cichlid diversification, and environmental variables for all lakes (Supplementary Information 2). We calculated maximum time for diversification for lineages using either the midpoint of geological age estimates for the lake or the mean stem age of the radiating group estimated from our calibrated molecular phylogenies.

To use the tree to account for phylogeny, we trimmed it to include only lineages that occur in lakes, and a single taxon to represent each diversifying lineage. For lineages present in several lakes, we added a tip to the tree for each instance the lineage had independently colonized a lake (see Supplementary Information 1.3), thereby accounting for each 'opportunity' for diversification.

We used phylogenetic logistic regression[28] to assess the relationship between each predictor variable and cichlid diversification state. Then, after assessing collinearity between predictor variables (Supplementary Information 3), we used multiple logistic regression to assess the combined influence of predictor variables on diversification state using a two-stage approach (Supplementary Information 5). First, we used AICc-based model averaging[19] to assess the relative importance of predictor variables. Second, we included predictor variables with relative importance values above 0.6 in phylogenetic logistic and phylogenetic hurdle Poisson regression models to attain phylogenetically corrected regression parameter estimates.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 17 January; accepted 16 April 2012.
Published online 10 June 2012.

1. Simpson, G. G. *The Major Features of Evolution* (Columbia Univ. Press, 1953).
2. Losos, J. B. Adaptive radiation, ecological opportunity, and evolutionary determinism. *Am. Nat.* **175,** 623–639 (2010).
3. Kisel, Y. & Barraclough, T. G. Speciation has a spatial scale that depends on levels of gene flow. *Am. Nat.* **175,** 316–334 (2010).
4. Schluter, D. *The Ecology of Adaptive Radiation* (Oxford Univ. Press, 2000).
5. Vamosi, S. M. The presence of other fish species affects speciation in threespine sticklebacks. *Evol. Ecol. Res.* **5,** 717–730 (2003).
6. MacArthur, R. H. & Wilson, E. O. Equilibrium-theory of insular zoogeography. *Evolution* 17, 373– (1963).
7. Price, S. A., Holzman, R., Near, T. J. & Wainwright, P. C. Coral reefs promote the evolution of morphological diversity and ecological novelty in labrid fishes. *Ecol. Lett.* **14,** 462–469 (2011).
8. Mittelbach, G. G. *et al.* Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. *Ecol. Lett.* **10,** 315–331 (2007).
9. Evans, K. L., Warren, P. H. & Gaston, K. J. Species-energy relationships at the macroecological scale: a review of the mechanisms. *Biol. Rev. Camb. Philos. Soc.* **80,** 1–25 (2005).
10. Kraaijeveld, K., Kraaijeveld-Smit, F. J. L. & Maan, M. E. Sexual selection and speciation: the comparative evidence revisited. *Biol. Rev. Camb. Philos. Soc.* **86,** 366–377 (2010).
11. Farrell, B. D. ''Inordinate fondness'' explained: why are there so many beetles? *Science* **281,** 555–559 (1998).
12. Liem, K. F. Evolutionary strategies and morphological innovations: cichlid pharyngeal jaws. *Syst. Zool.* **22,** 425–441 (1973).
13. Ricklefs, R. E. History and diversity: explorations at the intersection of ecology and evolution. *Am. Nat.* **170,** S56–S70 (2007).
14. Fryer, G. Some aspects of evolution in Lake Nyasa. *Evolution* **13,** 440–451 (1959).
15. Sturmbauer, C., Baric, S., Salzburger, W., Ruber, L. & Verheyen, E. Lake level fluctuations synchronize genetic divergences of cichlid fishes in African lakes. *Mol. Biol. Evol.* **18,** 144–154 (2001).
16. Seehausen, O. & van Alphen, J. M. Can sympatric speciation by disruptive sexual selection explain rapid evolution of cichlid diversity in Lake Victoria? *Ecol. Lett.* **2,** 262–271 (1999).
17. Seehausen, O. Evolution and ecological theory—chance, historical contingency and ecological determinism jointly determine the rate of adaptive radiation. *Heredity* **99,** 361–363 (2007).
18. Greenwood, P. H. Towards a phyletic classification of the 'genus' Haplochromis (Pisces, Cichlidae) and related taxa. Part 1. *Bull. Br. Mus. Nat. Hist. Zool.* **35,** 265–322 (1979).
19. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference* (Springer, 2002).
20. Ingram, T. Speciation along a depth gradient in a marine adaptive radiation. *Proc. R. Soc. B* **278,** 613–618 (2011).
21. Seehausen, O. *et al.* Speciation through sensory drive in cichlid fish. *Nature* **455,** 620–626 (2008).
22. Losos, J. B. & Schluter, D. Analysis of an evolutionary species–area relationship. *Nature* **408,** 847–850 (2000).
23. Lande, R. Rapid origin of sexual isolation and character divergence in a cline. *Evolution* **36,** 1–12 (1982).
24. Maan, M. E. & Seehausen, O. Ecology, sexual selection and speciation. *Ecol. Lett.* **14,** 591–602 (2011).
25. Trivers, R. L. in *Sexual Selection and the Descent of Man* (ed. Campbell, B.) 136–179 (Aldine, 1972).
26. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
27. Britton, T., Anderson, C. L., Jacquet, D., Lundqvist, S. & Bremer, K. Estimating divergence times in large phylogenetic trees. *Syst. Biol.* **56,** 741–752 (2007).
28. Ives, A. R. & Garland, T. Phylogenetic logistic regression for binary dependent variables. *Syst. Biol.* **59,** 9–26 (2010).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Ives, T. J. Davies and A. Mooers for analytical advice, P. McIntyre and C. Reidy Liermann for access to global energy data, S. Mwaiko, R. B. Stelkens, C. Katongo and U. Schliewen for unpublished DNA sequences, U. Schliewen, J. Jenson and O. Rittner for photographs, and A. McCune, D. Rabosky, R. Harrison, I. Lovette, E. Michel, G. Mittelbach, C. Melian, J. Brodersen, M. Maan, T. Ingram, B. Matthews, B. Dalziel, M. Pennell, J. Eastman, the Harmon laboratory group, the McCune laboratory group and the Seehausen laboratory group for discussions and comments on the manuscript. Bioinformatics facilities were supported by grants from the National Center for Research Resources (5P20RR016448-10) and the National Institute of General Medical Sciences (8 P20 GM103397-10) from the National Institutes of Health. This work was supported by the Swiss National Science Foundation project 31003A-118293 (to O.S.) and US National Science Foundation grant DEB 0919499 (to L.J.H.).

Author Contributions C.E.W., L.J.H. and O.S. designed the study. O.S. and C.E.W. collected the data. C.E.W. conducted the analyses. C.E.W., L.J.H. and O.S. wrote the paper.

Author Information Sequence data are deposited in the GenBank database under accession numbers listed in Supplementary Table 1. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to C.E.W. (catherine.wagner@eawag.ch) or O.S. (ole.seehausen@eawag.ch).

METHODS

Phylogenetic framework. We compiled sequence data for nine genes and 656 African cichlid species, with the goal of phylogenetically placing all African cichlid lineages present in lakes (Supplementary Information 1). The aligned, concatenated data set included a total of 6,947 base pairs.

We used a maximum likelihood approach in RAxML for phylogenetic analyses[26] (Supplementary Information 1.2). To account for phylogenetic uncertainty, we used 100 replicates of the rapid bootstrap algorithm in RAxML and estimated branch lengths for each of these bootstrap replicate topologies. To ultrametricize and time-calibrate this set of trees, we used PATHd8 (ref. 27). For time-calibration we used three geological dates and one fossil date: two dates associated with the breakup of Gondwana (the Africa–Madagascar split and the Madagascar–India split), the age of the earliest known fossil *Oreochromis*, and the age of Lake Nabugabo (Supplementary Information 1.2). We then drew 95% confidence intervals on node ages from the distribution of branching times estimated from this set of calibrated ultrametric trees.

Cichlid radiation data, ecological variables and species traits. We compiled information about presence of cichlid lineages in lakes across Africa, and the endemic diversity of the lineages present in each of these lakes (Supplementary Information 2). Because most colonizing lineages do not diversify, this data set is zero-inflated; the processes influencing radiation therefore should be analytically considered separately from the processes influencing the species richness of radiating lineages (see Supplementary Information 2.4). We thus coded each lineage in each lake as either 'diversifying' or 'non-diversifying' at two diversity thresholds. First, radiations with one or more intralacustrine speciation events (any lineage that had at least one endemic species in a lake co-occurring with its sister taxon, be it either a widespread species (in three cases) or a lake endemic itself); second, as a more conservative threshold for radiation, diversification events producing at least five endemic species. Analyses conducted at other thresholds produced qualitatively identical results to those at these threshold values. Single endemic species not co-occurring in the same lake with a sister taxon were not considered to be radiating lineages.

We compiled information about lineage-level character states for traits potentially linked to cichlid speciation, and environmental variables, and then used phylogenetic logistic regression[28] to analyse the association between these factors and diversification state. Tested lineage traits included the presence of a polygamous mating system, of mouthbrooding, of generalized egg dummies and specialized haplochromine-type egg dummies on the anal fin of male fish, and the presence of strong sexual dichromatism (Supplementary Information 2). Many of these traits have been proposed to be linked to sexual selection, and mouthbrooding has additionally been proposed as an ecological key innovation because it liberates cichlids from substrate-related habitat requirements for attachment and guarding of eggs[29]. These traits are rarely polymorphic within cichlid lineages. These few instances were coded by majority state, or as missing data (in one case, presence/absence of egg dummies in *Thoracochromis* of Lake Fwa, where majority state was ambiguous).

We compiled information on physical and environmental variables for all lakes in the data set. These included surface area, maximum depth, latitude, net solar radiation (the difference between the influx of solar energy and that reflected back into the atmosphere at a given geographic location, referred to simply as 'energy') and elevation (Supplementary Information 2). We chose these variables as the main factors correlating with lake type, habitat availability and climate that were available for many lakes. As a further environmental variable, we included the presence of large predatory fish (genera *Lates*, *Hydrocynus*, *Hepsetus*) because of their hypothesized influence on cichlid diversification[14,30].

We calculated maximum time for diversification for lineages using either the midpoint of geological age estimates for the lake (either basin age or most recent desiccation age) or the mean stem age of the radiating group estimated from our calibrated molecular phylogenies. We also conducted analyses using only geological lake ages, and these produced very similar results (Supplementary Tables 3 and 4).

Regression models. To account for phylogeny in regression models, we trimmed the best maximum likelihood topology to include only lineages that occur in lakes and a single taxon for each within-lake radiation. For lineages present in several lakes, we added a tip to the tree for each instance where the lineage is found in a unique lake, such that each lineage found in several lakes is represented as a polytomy with a tip corresponding to each lake where it occurs. We set branch lengths on these added tips to have a total length of that expected under a pure birth model (Supplementary Information 1.3). Using this approach, our trimmed and manipulated phylogenies had a branch for each 'opportunity' to diversify; that is, each instance a lineage entered a new lake.

We used phylogenetic logistic regression[28] to assess the relationship between single predictor variables and diversification state. To assess the combined influence of our predictor variables on cichlid diversification state, we used multiple logistic regression models. Before including the predictor variables in multiple regression models, we checked for collinearity between both continuous and binary predictor variables. We calculated Pearson correlation coefficients (r^2) for all pairs of continuous predictor variables. For binary predictor variables, we used the r^2 equivalent[31], r^2_L, as an assessment of collinearity (Supplementary Information 3). We removed one variable from each pair of predictor variables with r^2 (or r^2_L) greater than 0.7 after preliminary models including variables with higher correlations caused analytical problems (inflations of standard error, a diagnostic of collinearity problems in logistic regression[32]).

Because we discovered a strong correlation between lake depth and time for diversification during collinearity tests ($r^2 = 0.76$), we conducted further tests to determine the relative effects of time and depth. We excluded taxa from lakes greater than 150 m in depth ($n = 3$ of 46), leaving the remaining data subset uncorrelated in time and depth ($r^2 = 0.25$). We compared AIC values among models incorporating time, time + depth, and depth as predictors of cichlid diversification in this data set (Supplementary Information 3).

We examined the combined influence of predictor variables on diversification state in multiple regression models. Because likelihood-based phylogenetic logistic regression methods are not available, we used the following two-step approach. First, we used AICc-based model averaging[19] to evaluate the parameter estimates and the relative importance of predictor variables in a likelihood-based framework. We calculated model-averaged parameter estimates and standard errors for each predictor variable using relative AICc weights of models in which the variables appeared. We calculated the relative importance of each predictor variable as the sum of the AICc weights of all models that included this variable. Second, we included predictor variables with relative importance values above 0.6 in phylogenetic multiple logistic regression[28] models to attain phylogenetically corrected regression parameter estimates.

As an additional test of our results, we performed phylogenetic hurdle Poisson regression using the R package MCMCglmm[33], using the number of speciation events within each colonizing lineage as the response variable. This approach models two latent variables associated with the data: one associated with a binary process, the other modelling the non-zero response values in the data set as a Poisson process (Supplementary Information 5.3). We repeated these analyses over a set of 100 bootstrap replicate trees to account for phylogenetic uncertainty. Results were qualitatively identical to those from analyses modelling the binary process alone (Supplementary Information 5.3). Also using this modelling framework, we did post hoc tests for interaction effects between environmental and lineage-specific variables; these produced some evidence for interaction effects between lake depth and sexual dichromatism in predicting cichlid adaptive radiation (Supplementary Information 5.4).

29. Kuwamura, T. in *Fish communities in Lake Tanganyika* (eds Kawanabe, H., Hori, M. & Nagoshi, M.) 59–86 (Kyoto Univ. Press, 1997).
30. Worthington, E. B. & Ricardo, C. K. The fish of Lake Tanganyika (other than Cichlidae). *Proc. Zool. Soc. Lond.* **1936,** 1061–1112 (1936).
31. Menard, S. Coefficients of determination for multiple logistic regression analysis. *Am. Stat.* **54,** 17–24 (2000).
32. Quinn, G. P. & Keough, M. J. *Experimental Design and Data Analysis for Biologists* (Cambridge Univ. Press, 2002).
33. Hadfield, J. D. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* **33,** 1–22 (2010).

LETTER

doi:10.1038/nature11184

Proto-genes and *de novo* gene birth

Anne-Ruxandra Carvunis[1,2,3], Thomas Rolland[1,2], Ilan Wapinski[4], Michael A. Calderwood[1,2], Muhammed A. Yildirim[5], Nicolas Simonis[1,2]†, Benoit Charloteaux[1,2,6], César A. Hidalgo[7], Justin Barbette[1,2], Balaji Santhanam[1,2], Gloria A. Brar[8], Jonathan S. Weissman[8], Aviv Regev[9,10], Nicolas Thierry-Mieg[3], Michael E. Cusick[1,2] & Marc Vidal[1,2]

Novel protein-coding genes can arise either through re-organization of pre-existing genes or *de novo*[1,2]. Processes involving re-organization of pre-existing genes, notably after gene duplication, have been extensively described[1,2]. In contrast, *de novo* gene birth remains poorly understood, mainly because translation of sequences devoid of genes, or 'non-genic' sequences, is expected to produce insignificant polypeptides rather than proteins with specific biological functions[1,3–6]. Here we formalize an evolutionary model according to which functional genes evolve *de novo* through transitory proto-genes[4] generated by widespread translational activity in non-genic sequences. Testing this model at the genome scale in *Saccharomyces cerevisiae*, we detect translation of hundreds of short species-specific open reading frames (ORFs) located in non-genic sequences. These translation events seem to provide adaptive potential[7], as suggested by their differential regulation upon stress and by signatures of retention by natural selection. In line with our model, we establish that *S. cerevisiae* ORFs can be placed within an evolutionary continuum ranging from non-genic sequences to genes. We identify ~1,900 candidate proto-genes among *S. cerevisiae* ORFs and find that *de novo* gene birth from such a reservoir may be more prevalent than sporadic gene duplication. Our work illustrates that evolution exploits seemingly dispensable sequences to generate adaptive functional innovation.

Both genome-wide surveys and analyses of individual cases have shown that *de novo* gene birth has occurred throughout the evolution of many lineages, potentially affecting species-specific adaptations and evolutionary radiations[1,2,5,6,8,9]. Genes are thought to emerge *de novo* when non-genic sequences become transcribed, acquire ORFs and the corresponding non-genic transcripts access the translation machinery[1,2,4,5,8]. However, it is hard to reconcile this proposed mechanism with expectations that non-genic sequences should lack translational activity and, even if translated, should encode insignificant polypeptides[1,3,4,6]. Evidence of associations between non-genic transcripts and ribosomes has suggested that non-genic sequences may occasionally be translated, which could provide raw material for natural selection[6]. It has also been speculated that genes that originate *de novo* could initially be simple and gradually become more complex over evolutionary time[4]. These ideas are consistent with reports showing that genes that emerged recently are shorter, less expressed and more rapidly diverging than other genes[1,10–13]. We developed an integrative evolutionary model whereby *de novo* gene birth proceeds through intermediate and reversible proto-gene stages, mirroring the well-described pseudo-gene stages of gene death (Fig. 1a)[14].

We investigated this model at genome scale in the context of *de novo* gene birth in *S. cerevisiae*[8,10]. In *S. cerevisiae*, a minimal length threshold of 300 nucleotides was originally used to delineate ORFs likely to be genes from non-genic ORFs occurring by chance in non-genic sequences[15]. The resulting gene catalogue has undergone numerous adjustments[16], with currently ~6,000 ORFs annotated as genes and ~261,000 unannotated ORFs containing at least three codons considered to be non-genic ORFs (Supplementary Fig. 1). Non-genic sequences are broadly transcribed in *S. cerevisiae*[17], their overexpression is mostly non-toxic[18], and the corresponding transcripts can associate with ribosomes, often at AUG start codons[6,19]. We reasoned that translation of non-genic ORFs could be more common than expected. Such translation events would not systematically lead to *de novo* gene birth, as the corresponding polypeptides would not necessarily have specific biological functions. Instead, upon translation, non-genic ORFs would become proto-genes (Fig. 1b). Proto-genes would provide adaptive potential[6] by exposing genetic variations that are usually hidden in non-genic sequences. A subset of proto-genes could occasionally be retained over evolutionary time, for instance if providing an advantage to the organism under specific environmental conditions. Retained proto-genes could gradually evolve the characteristics of genes, whereas other proto-genes might lose the ability to be translated. Such a reservoir of proto-genes would allow evolutionary innovations to be attempted without affecting existing genes.

This evolutionary model leads to the following predictions: (1) the structural and functional characteristics of *S. cerevisiae* ORFs (for example, length, expression level or sequence composition) should reflect an evolutionary continuum ranging from non-genic ORFs to genes; (2) many non-genic ORFs should be translated; and (3) ORFs that emerged recently should occasionally have adaptive functions retained by natural selection.

To examine these predictions, we estimated the order of emergence of *S. cerevisiae* ORFs (Fig. 1c). Annotated ORFs were classified into ten groups based on their conservation throughout the Ascomycota phylogeny (Supplementary Fig. 2). Of ~6,000 annotated ORFs, ~2% are found only in *S. cerevisiae* ($ORFs_1$) (Supplementary Fig. 2)[10] and ~12% are found only in the four closely related *Saccharomyces sensu stricto* species ($ORFs_{1-4}$). The ~88% of annotated ORFs found outside of this group ($ORFs_{5-10}$) are well characterized and can confidently be considered genes. $ORFs_{1-4}$ are poorly characterized and their annotation as genes is debatable (Supplementary Fig. 2)[16,20]. The weak conservation of $ORFs_{1-4}$ suggests that they emerged recently, which we corroborated using gene duplication events to control for relative time of emergence (Supplementary Fig. 3). We estimate that over 97% of $ORFs_{1-4}$ originated *de novo* rather than by cross-species transfer, which could also explain their weak conservation (Supplementary Information). $ORFs_{1-4}$ often partially overlap $ORFs_{5-10}$, which seems incompatible with cross-species transfer, or preferentially lie within

[1]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [3]UJF-Grenoble 1/CNRS/TIMC-IMAG UMR 5525, Computational and Mathematical Biology Group, Grenoble F-38041, France. [4]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115, USA. [5]Center for International Development and Harvard University, Cambridge, Massachusetts 02138, USA. [6]Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liege, 4000 Liege, Wallonia-Brussels Federation, Belgium. [7]The MIT Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [8]Howard Hughes Medical Institute, Department of Cellular and Molecular Pharmacology, University of California, San Francisco, and California Institute for Quantitative Biosciences, San Francisco, California 94158, USA. [9]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [10]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. †Present address: Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe), Campus Plaine, Free University of Brussels, 1050 Brussels, Wallonia-Brussels Federation, Belgium.

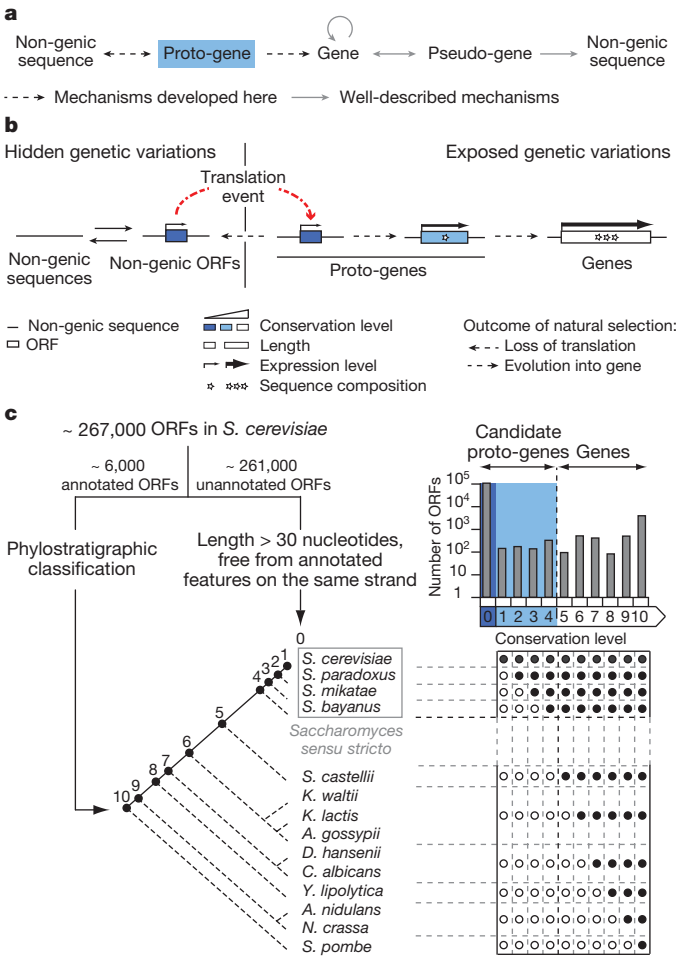


Figure 1 | From non-genic sequences to genes through proto-genes. **a**, Proto-genes mirror for gene birth the well-described pseudo-genes for gene death. Circular arrow indicates gene origination from pre-existing genes, such as through gene duplication. Pseudo-genes are highly related to existing genes but have accumulated disabling mutations and translation of functional proteins is no longer possible[14]. The premise that pseudo-gene formation represents irreversible gene death has been challenged by reports of pseudo-gene resurrection[14] (bidirectional arrow). After enough evolutionary time pseudo-gene decay renders them indistinguishable from non-genic sequences (unidirectional arrow). Whereas pseudo-genes resemble known genes, proto-genes resemble no known genes. Proto-genes arise in non-genic sequences and either revert to non-genic sequences or evolve into genes (bidirectional arrow). There can be no reversion of genes to proto-genes (unidirectional arrow) as gene decay engenders pseudo-genes. **b**, Details of the proposed model for the gradual emergence of protein-coding genes in non-genic sequences via proto-genes. Solid arrows indicate the reversible emergence of ORFs in non-genic transcripts, or of transcripts containing non-genic ORFs. Examples where transcript appearance precedes ORF appearance have been described[1,2,8], but the reverse order of events cannot be ruled out. Arrows representing expression level symbolize transcription (hidden genetic variation) or transcription and translation (exposed genetic variation). The variations in width of these arrows reflect changes in expression level resulting, at least in part, from changes in regulatory sequences. Sequence composition refers to codon usage, amino acid abundances and structural features. **c**, Assigning conservation levels to *S. cerevisiae* ORFs. Conservation levels of annotated ORFs were assigned according to comparisons along the reconstructed phylogenetic tree, by inferring their presence (filled circles) or absence (open circles) in the different species according to the phylostratigraphy principle (Supplementary Information)[1]. Top right, number of ORFs assigned to each conservation level (logarithmic scale). *A. gossypii*, *Ashbya* (*Eremothecium*) *gossypii*; *A. nidulans*, *Aspergillus nidulans*; *C. albicans*, *Candida albicans*; *D. hansenii*, *Debaryomyces hansenii*; *K. lactis*, *Kluyveromyces lactis*; *K. waltii*, *Kluyveromyces* (*Lachancea*) *waltii*; *N. crassa*, *Neurospora crassa*; *S. pombe*, *Schizosaccharomyces pombe*.

subtelomeric regions whose instability may facilitate *de novo* emergence (Supplementary Fig. 4). In addition to classifying $ORFs_{1-10}$, we assigned a conservation level of 0 to ~108,000 unannotated ORFs longer than 30 nucleotides and free from overlap with annotated features on the same strand ($ORFs_0$) (Supplementary Information). $ORFs_0$ and $ORFs_{1-4}$ constituted our initial list of candidate proto-genes.

To test the evolutionary continuum prediction, we first verified that ORF conservation level correlates positively with length and expression level (Fig. 2a and Supplementary Fig. 5)[1,10–12]. These correlations suggest that genes evolve from non-genic ORFs that lengthen and increase in expression level over evolutionary time. A negative correlation between ORF length and expression level[21] was observed among

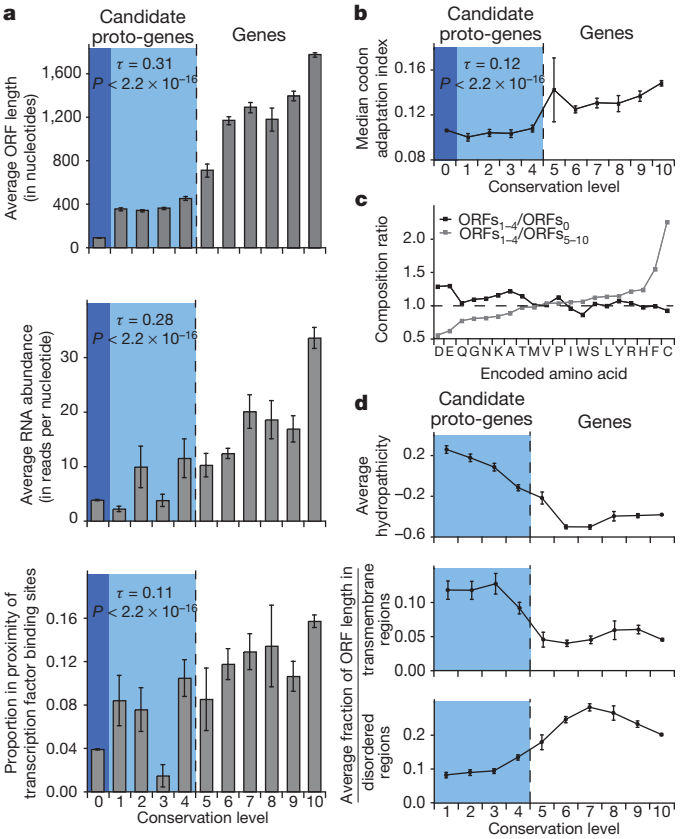


Figure 2 | Existence of an evolutionary continuum ranging from non-genic ORFs to genes through proto-genes. **a**, Length (top; error bars represent s.e.m.), RNA expression level (middle; error bars represent s.e.m.), and proximity to transcription factor binding sites (bottom; error bars represent standard error of the proportion) of ORFs correlate with conservation level (Supplementary Table 4). *P* and τ, Kendall's correlation statistics. Estimation of RNA abundance from RNA-Seq[25] in rich conditions. The positive correlation between proximity to transcription factor binding sites and conservation level is shown for a window of 200 nucleotides and holds when considering windows of 300, 400 and 500 nucleotides (Kendall's $\tau = 0.14, 0.16, 0.17$, respectively; $P < 2.2 \times 10^{-16}$ in each case). **b**, Codon bias increases with conservation level (Supplementary Table 4). Codon bias estimated using the codon adaptation index (Supplementary Information). *P* and τ, Kendall's correlation statistics. Error bars represent s.e.m. The large s.e.m. observed for $ORFs_5$ may be related to the whole genome duplication event (Supplementary Fig. 3). **c**, Relative amino acid abundances shift with increasing conservation level. For each encoded amino acid, the ratio between its frequency in $ORFs_{1-4}$ and its frequency in $ORFs_{5-10}$ (grey), or the ratio between its frequency in $ORFs_{1-4}$ and its frequency in $ORFs_0$ (black), is plotted. Enrichment of cysteine in proteins encoded by $ORFs_{1-4}$ relative to those encoded by $ORFs_{5-10}$ ($P < 1.8 \times 10^{-150}$, hypergeometric test) corresponds to 3.6 ± 0.1 residues (mean, s.e.m.) per translation product. **d**, Predicted structural features of ORF translation products correlate with conservation level. $ORFs_0$ were not included in these analyses as their short length hinders the reliability of structural predictions. Error bars represent s.e.m.

ORFs$_{5-10}$, but not among ORFs$_{1-4}$ (Supplementary Fig. 5). Thus, some ORFs may increase in expression level at different rates than they increase in length over evolutionary time. Lengthening of ORFs could occur by loss of stop codons, possibly following translational read-through, by shift of start codons or by duplication followed by fusion with other ORFs[10,22]. Increase in ORF expression level could be mediated by recruitment of existing regulatory elements[1]. The proportion of ORFs located in the vicinity of transcription factor binding sites increases with conservation level, suggesting that novel regulatory elements could also emerge (Fig. 2a)[1].

In line with a study of codon evolution in metazoans[23], we observed a positive correlation between codon usage bias and conservation level (Fig. 2b). Relative abundances of amino acids in proteins encoded by ORFs$_{1-4}$ show levels intermediate between those in proteins encoded by ORFs$_{5-10}$ and in hypothetical translation products of ORFs$_0$ (Fig. 2c), similar to observations in bacteria[24]. Probably owing to this biased sequence composition, ORFs$_{1-4}$ exhibit a higher hydropathicity, a higher tendency to form transmembrane regions and a lower propensity for intrinsic structural disorder[10] than ORFs$_{5-10}$ (Fig. 2d). Taken together, our observations support the existence of an evolutionary continuum ranging from non-genic ORFs to genes.

To assess the extent of non-genic translation, we searched for signatures of translation of ORFs$_0$ at genome scale in a ribosome footprinting data set generated in both rich and starvation conditions[25]. In this data set, ~1% of sequencing reads could not be mapped to ORFs$_{1-10}$. We developed a stringent pipeline to detect unequivocal translation signatures for ORFs$_0$ located on transcripts associated with ribosomes (Fig. 3a and Supplementary Fig. 6). We found that 1,139 of ~108,000 ORFs$_0$ show such evidence of translation (ORFs$_0^+$). This number is significantly higher than expected if the ribosome footprinting assay was non-specific, or if the presence of ribosomes on non-genic transcripts was unrelated to the presence of ORFs$_0$ (Fig. 3b). These ORFs$_0^+$ are enriched in adenine at position −3 from the start codon, which probably favours translation initiation (Fig. 3c and Supplementary Information). We verified that ORFs$_0^+$ did not originate from gene duplication or cross-species transfer and are not genes that have failed to be annotated due to their short length (Supplementary Information). The 1,139 ORFs$_0^+$ therefore appear to be translated non-genic ORFs.

We detected strong differential translation of ORFs$_0^+$ and ORFs$_{1-4}$ in starvation or rich conditions, whereas most ORFs$_{5-10}$ are translated in both conditions (Fig. 3d and Supplementary Fig. 6). We found that the binding sites of four transcription factors involved in mating and stress response are preferentially located close to ORFs$_0^+$ and ORFs$_{1-4}$ (Supplementary Table 1) and that ORFs$_{1-4}$ are enriched in the Gene Ontology term "response to stress" (Supplementary Table 2). Recently emerged ORFs may provide adaptive functions in response to environmental stress.

Retention by natural selection was measured by comparing the genome sequences of eight *S. cerevisiae* strains to evaluate the tendency of ORF sequences to be purged of non-synonymous mutations (purifying selection) relative to expectations under neutral evolution. Most ORFs$_0^+$ and ORFs$_{1-4}$ do not exhibit a significant deviation from neutral evolution, yet ~3% of ORFs$_0^+$ and 9–25% of ORFs$_{1-4}$ appear under purifying selection (Fig. 3e). This fraction increases with conservation level, in line with the proposed evolutionary continuum (Supplementary Fig. 7 and Supplementary Information). Our observations suggest that recently emerged ORFs occasionally acquire adaptive functions that are retained by natural selection, in agreement with findings in primates and with evolutionary models derived from inter-species comparisons[12,13,26].

Overall, our results show that *de novo* gene birth could proceed through proto-genes. From the initial comprehensive set of candidate proto-genes (all ORFs$_0$ and ORFs$_{1-4}$), we excluded ORFs$_0$ that seem to lack translation signatures according to our stringent pipeline (Supplementary Fig. 6). The 25 ORFs$_4$ that are longer than 300 nucleotides, show signatures of translation and are under purifying selection, can confidently be considered genes despite being weakly conserved (Fig. 4a and Supplementary Fig. 8). The remaining 1,891

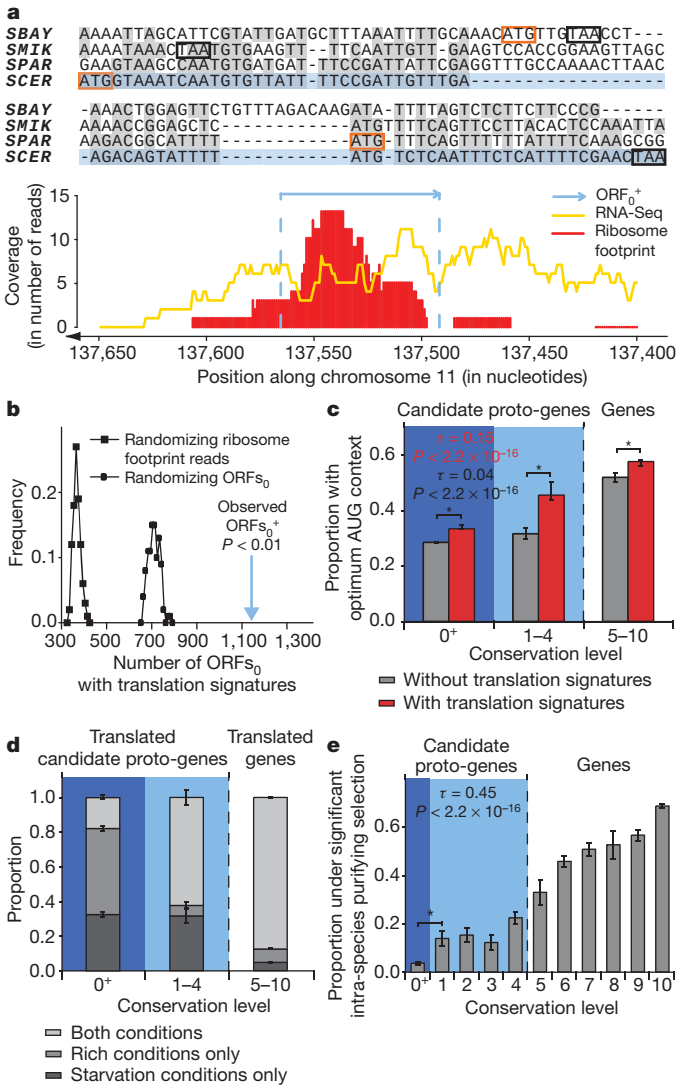


Figure 3 | Translation and adaptive potential of recently emerged ORFs. **a**, Example of an ORF$_0^+$ showing signatures of translation in starvation conditions. Syntenic regions in *Saccharomyces sensu stricto* species are aligned. Orange and black boxes indicate in-frame start and stop sites, respectively. *SCER*, *S. cerevisiae*; *SPAR*, *S. paradoxus*; *SMIK*, *S. mikatae*; *SBAY*, *S. bayanus*. **b**, Significance of the observed number of ORFs$_0^+$. Distribution of the number of ORFs$_0$ expected to show signatures of translation if the ribosome footprinting assay were non-specific (as modelled by randomizing footprint reads positions 100 times; squares), or if the presence of ribosomes on non-genic transcripts were not related to the presence of ORFs$_0$ (as modelled by randomizing ORFs$_0$ positions 100 times; circles). *P*, empirical *P* value. **c**, AUG context of ORFs with and without translation signatures. The presence of an adenine at position −3 from the start codon indicates optimum AUG context (Supplementary Information). *P* and τ, Kendall's correlation statistics. Asterisks mark significant differences between ORFs with and without translation signatures ($P < 0.05$, Fisher's exact test). **d**, Candidate proto-genes tend to undergo condition-specific translation. **e**, Signatures of intra-species purifying selection. The positive correlation (Supplementary Table 4) holds when only considering ORFs that are free from overlap with ORFs$_{1-10}$ (Supplementary Fig. 7), and is not entirely driven by the interdependence between strength of purifying selection and expression level (Supplementary Information)[29,30]. Asterisk marks a significant difference in proportion of ORFs under significant intra-species purifying selection between ORFs$_0^+$ and ORFs$_1$ ($P = 0.0001$, hypergeometric test). *P* and τ, Kendall's correlation statistics. Error bars represent standard error of the proportion in all panels.

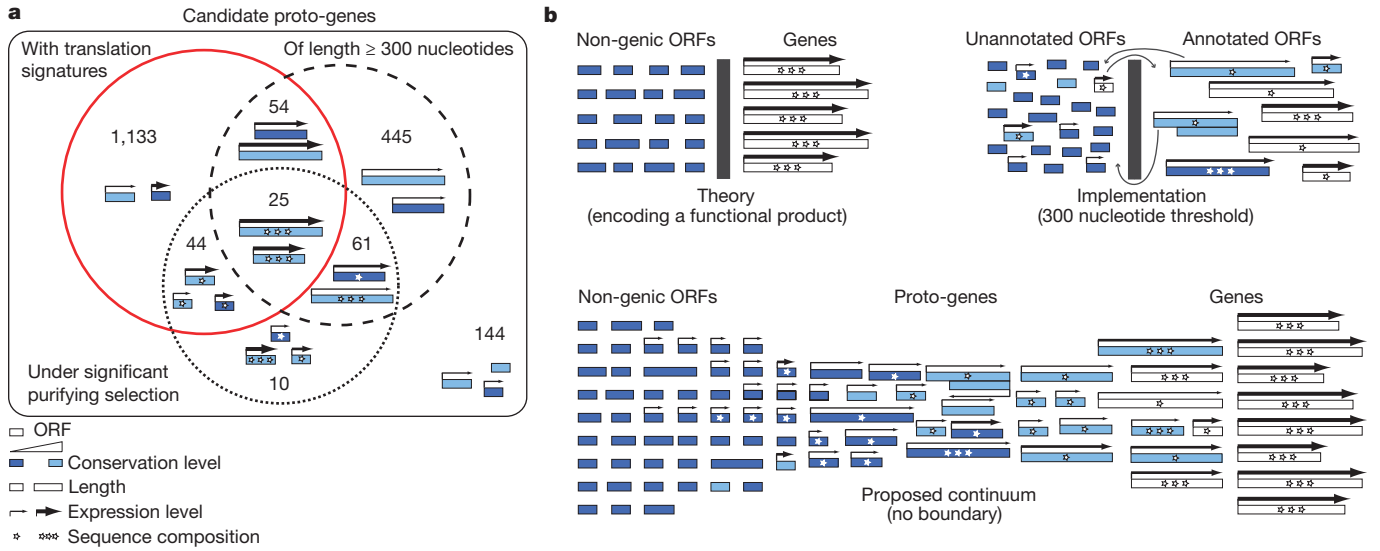


Figure 4 | Identification of proto-genes in a continuum ranging from non-genic ORFs to genes. **a**, Characterization of candidate proto-genes ($ORFs_0^+$ and $ORFs_{1-4}$). Venn diagram not drawn to scale. **b**, The binary model of annotation (top) and the proposed continuum (bottom).

ORFs (1,139 $ORFs_0^+$ and 752 $ORFs_{1-4}$) present characteristics intermediate between non-genic ORFs and genes, meeting our proto-gene designation (Supplementary Table 3). We propose to place these ORFs in a continuum where strict annotation boundaries no longer have to be set (Fig. 4b).

Gene birth mechanisms involving re-organization of pre-existing genes, notably following gene duplication, have long been regarded as the predominant source of evolutionary innovation[1,2]. Since the split between *S. cerevisiae* and *S. paradoxus*, sporadic gene duplications have generated between one and five novel genes[27]. In contrast, 19 of the 143 $ORFs_1$ that arose *de novo* during the same evolutionary period were found under purifying selection. Therefore, *de novo* gene birth seems to be more prevalent than previously supposed[3,10,12], in agreement with recent estimations in humans and other primates[1,9]. The involvement of proto-genes in *de novo* emergence of protein-coding genes in *S. cerevisiae* probably holds for other species and may extend to RNA genes and regulatory elements. Examination of translation program remodelling upon stress, in light of our evolutionary model, may further understanding of phenotypic diversity and plasticity of cellular systems[7,28].

METHODS SUMMARY

Detection of translation signatures. The mapping of ribosome footprint reads to ORFs does not necessarily indicate full-length, ORF-specific translation events[6,25]. To model the number of $ORFs_0^+$ expected if the detected presence of ribosomes on non-genic sequences was not related to the presence of $ORFs_0$, we randomized the positions of $ORFs_0$ while maintaining their length distribution and the observed positions of RNA-Seq and footprint reads. To model the number of $ORFs_0^+$ expected if footprint reads observed outside of annotated ORFs were non-specific, we randomized the positions of footprint reads throughout non-genic sequences while maintaining the length distribution of footprint reads, the positions of RNA-Seq reads and the positions of $ORFs_0$. We optimized three parameters with regard to these two null models: (1) the proportion of ORF length covered in RNA-Seq and footprint reads was fixed at 50% minimum; (2) the factor by which the number of footprint reads per nucleotide in the ORF should be higher than the number of footprint reads per nucleotide in surrounding up- and downstream windows was fixed at a minimum of 5; and (3) the size of these windows was fixed at 300 nucleotides. Any two $ORFs_0$ that partially overlap on the same strand and show translation signatures in the same experimental conditions were both eliminated from the set of $ORFs_0$ considered to show translation signatures.
Significant purifying selection signatures. We estimated the number of synonymous mutations per synonymous site (dS) and the number of non-synonymous mutations per non-synonymous site (dN) for each ORF present without disruptive mutations in eight *S. cerevisiae* strains. The likelihood of the dN/dS ratio was determined under two distinct null models: assuming neutral evolution (the rates of synonymous and non-synonymous substitutions are equal) and not assuming neutral evolution. All ORFs with $dN/dS < 1$ and $P < 0.05$ (chi-squared distribution of likelihoods with one degree of freedom) were considered to be subject to significant purifying selection.

Received 2 November 2011; accepted 8 May 2012.
Published online 24 June 2012.

1. Tautz, D. & Domazet-Loso, T. The evolutionary origin of orphan genes. *Nature Rev. Genet.* **12,** 692–702 (2011).
2. Kaessmann, H. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* **20,** 1313–1326 (2010).
3. Jacob, F. Evolution and tinkering. *Science* **196,** 1161–1166 (1977).
4. Siepel, A. Darwinian alchemy: human genes from noncoding DNA. *Genome Res.* **19,** 1693–1695 (2009).
5. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. & Bosch, T. C. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet.* **25,** 404–413 (2009).
6. Wilson, B. A. & Masel, J. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol. Evol.* **3,** 1245–1252 (2011).
7. Jarosz, D. F., Taipale, M. & Lindquist, S. Protein homeostasis and the phenotypic manifestation of genetic diversity: principles and mechanisms. *Annu. Rev. Genet.* **44,** 189–216 (2010).
8. Cai, J., Zhao, R., Jiang, H. & Wang, W. *De novo* origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* **179,** 487–496 (2008).
9. Wu, D. D., Irwin, D. M. & Zhang, Y. P. *De novo* origin of human protein-coding genes. *PLoS Genet.* **7,** e1002379 (2011).
10. Ekman, D. & Elofsson, A. Identifying and quantifying orphan protein sequences in fungi. *J. Mol. Biol.* **396,** 396–405 (2010).
11. Lipman, D. J., Souvorov, A., Koonin, E. V., Panchenko, A. R. & Tatusova, T. A. The relationship of protein conservation and sequence length. *BMC Evol. Biol.* **2,** 20 (2002).
12. Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V. & Lipman, D. J. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc. Natl Acad. Sci. USA* **106,** 7273–7280 (2009).
13. Cai, J. J. & Petrov, D. A. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol. Evol.* **2,** 393–409 (2010).
14. Zheng, D. & Gerstein, M. B. The ambiguous boundary between genes and pseudogenes: the dead rise up, or do they? *Trends Genet.* **23,** 219–224 (2007).
15. Oliver, S. G. *et al.* The complete DNA sequence of yeast chromosome III. *Nature* **357,** 38–46 (1992).
16. Fisk, D. G. *et al. Saccharomyces cerevisiae* S288C genome annotation: a working hypothesis. *Yeast* **23,** 857–865 (2006).
17. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–1349 (2008).
18. Boyer, J. *et al.* Large-scale exploration of growth inhibition caused by overexpression of genomic fragments in *Saccharomyces cerevisiae*. *Genome Biol.* **5,** R72 (2004).
19. Brar, G. A. *et al.* High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* **335,** 552–557 (2012).
20. Li, Q. R. *et al.* Revisiting the *Saccharomyces cerevisiae* predicted ORFeome. *Genome Res.* **18,** 1294–1303 (2008).
21. Jansen, R. & Gerstein, M. Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Res.* **28,** 1481–1488 (2000).

22. Giacomelli, M. G., Hancock, A. S. & Masel, J. The conversion of 3′ UTRs into coding regions. *Mol. Biol. Evol.* **24,** 457–464 (2007).
23. Prat, Y., Fromer, M., Linial, N. & Linial, M. Codon usage is associated with the evolutionary age of genes in metazoan genomes. *BMC Evol. Biol.* **9,** 285 (2009).
24. Yomtovian, I., Teerakulkittipong, N., Lee, B., Moult, J. & Unger, R. Composition bias and the origin of ORFan genes. *Bioinformatics* **26,** 996–999 (2010).
25. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–223 (2009).
26. Vishnoi, A., Kryazhimskiy, S., Bazykin, G. A., Hannenhalli, S. & Plotkin, J. B. Young proteins experience more variable selection pressures than old proteins. *Genome Res.* **20,** 1574–1581 (2010).
27. Gao, L. Z. & Innan, H. Very low gene duplication rate in the yeast genome. *Science* **306,** 1367–1370 (2004).
28. Hayden, E. J., Ferrada, E. & Wagner, A. Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature* **474,** 92–95 (2011).
29. Pal, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* **158,** 927–931 (2001).
30. Drummond, D. A., Raval, A. & Wilke, C. O. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23,** 327–337 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank L. Duret, E. Levy, J. Vandenhaute, Q. Li, H. Yu, P. Braun, M. Dreze, C. Foo, M. Mann, N. Kulak, J. Cox, C. Maire and S. Jhavery-Schneider as well as members of the Center for Cancer Systems Biology (CCSB), in particular A. Dricot-Ziter, A. MacWilliams, F. Roth, Y. Jacob and D. Hill for discussions and proofreading. A.R. was supported by a National Institute of Health Pioneer Award, a Career Award at the Scientific Interface from the Burroughs Wellcome Fund and the Howard Hughes Medical Institute (HHMI). I.W. is a HHMI fellow of the Damon Runyon Cancer Research Institute. G.A.B. was supported by American Cancer Society Postdoctoral fellowship 117945-PF-09-136-01-RMC. M.V. is a Chercheur Qualifié Honoraire from the Fonds de la Recherche Scientifique (FRS-FNRS, Wallonia-Brussels Federation, Belgium). This work was supported by the grant R01-HG006061 from the National Human Genome Research Institute awarded to M.V.

Author Contributions A.-R.C., I.W., M.E.C. and M.V. conceived the project. A.-R.C. led the project and performed most of the analyses. T.R. evaluated cross-species transfer events, optimized the ribosome footprint analysis pipeline and assisted in other analyses. I.W. designed the conservation level tool and calculated most of the purifying selection statistics. M.A.C., C.A.H., A.R. and N.T.-M. advised on the research. M.A.Y. aligned the sequencing reads. B.S. predicted disordered and transmembrane regions and assisted in the cross-species transfer analyses. N.S. and B.C. assisted in analyses. G.A.B. and J.S.W. shared their expertise in ribosome footprinting data analysis and provided the meiosis ribosome footprinting raw and processed data. A.-R.C., T.R., M.E.C. and M.V. designed the figures. All authors contributed to writing the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.V. (marc_vidal@dfci.harvard.edu).

LETTER

doi:10.1038/nature11174

Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing

Magnus Manske[1,2]*, Olivo Miotto[2,3]*, Susana Campino[1,2], Sarah Auburn[1,2,4], Jacob Almagro-Garcia[1,2,5], Gareth Maslen[1,2], Jack O'Brien[2,5], Abdoulaye Djimde[6], Ogobara Doumbo[6], Issaka Zongo[7], Jean-Bosco Ouedraogo[7], Pascal Michon[8], Ivo Mueller[8], Peter Siba[8], Alexis Nzila[9], Steffen Borrmann[9], Steven M. Kiara[9], Kevin Marsh[9], Hongying Jiang[10], Xin-Zhuan Su[10], Chanaki Amaratunga[10], Rick Fairhurst[10], Duong Socheat[11], Francois Nosten[3,12,13], Mallika Imwong[14], Nicholas J. White[3,13], Mandy Sanders[1], Elisa Anastasi[1], Dan Alcock[1], Eleanor Drury[1], Samuel Oyola[1], Michael A. Quail[1], Daniel J. Turner[1], Valentin Ruano-Rubio[1,2,5], Dushyanth Jyothi[1,2], Lucas Amenga-Etego[2,5,15], Christina Hubbart[5], Anna Jeffreys[5], Kate Rowlands[5], Colin Sutherland[16], Cally Roper[16], Valentina Mangano[17], David Modiano[17], John C. Tan[18], Michael T. Ferdig[18], Alfred Amambua-Ngwa[19], David J. Conway[16,19], Shannon Takala-Harrison[20], Christopher V. Plowe[20], Julian C. Rayner[1], Kirk A. Rockett[1,2,5], Taane G. Clark[1,2,16], Chris I. Newbold[1,2,21], Matthew Berriman[1], Bronwyn MacInnis[1,2] & Dominic P. Kwiatkowski[1,2,5]

Malaria elimination strategies require surveillance of the parasite population for genetic changes that demand a public health response, such as new forms of drug resistance[1,2]. Here we describe methods for the large-scale analysis of genetic variation in *Plasmodium falciparum* by deep sequencing of parasite DNA obtained from the blood of patients with malaria, either directly or after short-term culture. Analysis of 86,158 exonic single nucleotide polymorphisms that passed genotyping quality control in 227 samples from Africa, Asia and Oceania provides genome-wide estimates of allele frequency distribution, population structure and linkage disequilibrium. By comparing the genetic diversity of individual infections with that of the local parasite population, we derive a metric of within-host diversity that is related to the level of inbreeding in the population. An open-access web application has been established for the exploration of regional differences in allele frequency and of highly differentiated loci in the *P. falciparum* genome.

The genetic diversity and evolutionary plasticity of *P. falciparum* are major obstacles for malaria elimination. New forms of resistance against antimalarial drugs are continually emerging[1,2], and new forms of antigenic variation are a critical point of vulnerability for future malaria vaccines. Effective tools are needed to detect evolutionary changes in the parasite population and to monitor the spread of genetic variants that affect malaria control.

Here we describe the use of deep sequencing to analyse *P. falciparum* diversity, using blood samples from patients with malaria. The *P. falciparum* genome has several unusual features that greatly complicate sequence analysis, such as extreme AT bias, large tracts of non-unique sequence and several large families of intensely polymorphic genes[3]. Our aim was therefore not to determine the entire genome sequence of individual field samples—which would be prohibitively expensive with current technologies—but to define an initial set of single nucleotide polymorphisms (SNPs) distributed across the *P. falciparum* genome, whose genotype can be ascertained with confidence in parasitized blood samples by deep sequencing.

An additional complication in the analysis of *P. falciparum* genome variation is that the billions of haploid parasites that infect a single individual can be a complex mixture of genetic types. Previous studies[4–8] have largely focused on laboratory-adapted parasite clones, but the within-host diversity of natural infections is of fundamental biological interest. Parasites in the blood replicate asexually, but when they are taken up in the blood meal of an *Anopheles* mosquito they undergo sexual mating. If the parasites in the blood are of diverse genetic types, this process of sexual mating can generate novel recombinant forms. Deep sequencing provides new ways of investigating within-host diversity and the role of sexual recombination in parasite evolution.

P. falciparum DNA was obtained from blood samples collected from 290 patients with malaria at clinics in Burkina Faso, Cambodia, Kenya, Mali, Papua New Guinea and Thailand (Supplementary Table 1). For 149 samples we used the conventional method of growing the parasites in short-term blood culture before extracting the *P. falciparum* DNA. For 141 samples we used a new method by which *P. falciparum* DNA is extracted directly from venous blood samples after the removal of leukocytes[9]. We refer to these as cultured and direct samples, respectively.

Paired-end sequence reads were generated (median 7×10^8 base pairs per sample) by using the Illumina Genome Analyzer platform. Sequence analysis was divided into stages of SNP discovery, quality control filtering, genotyping and validation (see Supplementary Methods and Supplementary Fig. 1). After alignment to the 3D7 reference genome[3], non-coding regions had a much lower read depth than coding regions (Supplementary Fig. 2): this can be ascribed to their high AT content (non-coding 87% AT, coding 70% AT). Read depth was also low in the highly polymorphic *var*, *rifin* and *stevor* coding regions (Supplementary Fig. 3). For the purposes of this study, to decrease genotyping errors due to low coverage or copy number variation we excluded all non-coding regions, as well as coding regions at the extremes of the read depth distribution. After these exclusions we were left with 70% of all exonic positions across the genome, with

[1]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK. [2]MRC Centre for Genomics and Global Health, University of Oxford, Oxford OX3 7BN, UK. [3]Mahidol-Oxford Tropical Medicine Research Unit, Mahidol University, Bangkok 10400, Thailand. [4]Menzies School of Health Research, Charles Darwin University, Darwin, Northern Territories 0811, Australia. [5]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [6]Malaria Research and Training Centre, Faculty of Medicine, University of Bamako, Bamako, Mali. [7]Institut de Recherche en Sciences de la Santé, Direction Régionale de l'Ouést, Bobo-Dioulasso, Burkina Faso. [8]Papua New Guinea Institute of Medical Research, Madang 511, Papua New Guinea. [9]KEMRI/Wellcome Trust Research Program, Kilifi, Kenya. [10]National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, Maryland 20892, USA. [11]Cambodia National Malaria Centre, Phnom Penh, Cambodia. [12]Shoklo Malaria Research Unit, Mae Sot, Tak 63110, Thailand. [13]Centre for Tropical Medicine, University of Oxford, Oxford OX3 7LJ, UK. [14]Department of Molecular Tropical Medicine and Genetics, Faculty of Tropical Medicine, Mahidol University, Bangkok 10400, Thailand. [15]Navrongo Health Centre, Navrongo, Ghana. [16]London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK. [17]Department of Public Health Sciences, University of Rome 'La Sapienza', Rome 00185, Italy. [18]The Eck Institute for Global Health, Department of Biological Sciences, University of Notre Dame, Notre Dame, Indiana 4655, USA. [19]MRC Laboratories, Fajara, The Gambia. [20]Centre for Vaccine Development, University of Maryland, Baltimore, Maryland 21201, USA. [21]Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK.
*These authors made equal contributions to this work.

more than 50% of exonic positions for 71% of genes, and more than 70% for 54% of genes (Supplementary Table 2).

Within-host diversity complicates the process of excluding sequencing and alignment errors that are manifested as false heterozygous genotypes. Two approaches were identified to address this problem (see Supplementary Methods). We scored each position in the reference genome for its degree of uniqueness, and this was found to be a strong predictor of false heterozygous genotypes. We also observed a relationship between the population allele frequency of a SNP and its average level of within-sample heterozygosity, analogous to the Hardy–Weinberg relationship in diploid organisms. This enabled us to exclude SNPs that had excessive levels of within-sample heterozygosity relative to their population frequency.

After applying the above filters, and excluding SNPs and samples with high levels of missing data, we obtained a final data set of 86,158 SNPs genotyped in 227 samples (120 direct and 107 cultured) in which a median of 98% samples had valid genotyping data for each SNP, and a median of 98% SNPs had valid genotyping data for each sample (Supplementary Fig. 4). This set of 86,158 SNPs (here referred to as the 86k SNP set) represents 10% of the SNPs discovered at the initial stage of sequence alignment. Comparison with the PlasmoDB 5.5 database indicates that 77,283 (89%) of these SNPs are novel, but it should be noted that previous genome-wide SNP discovery efforts have largely been based on low-coverage capillary sequencing, and the overall error rate is unknown[4–6].

The accuracy of genotype calls in the 86k SNP set was evaluated by five independent approaches (see Supplementary Methods). We examined the evidence for 275 putative novel SNPs using independent data from PCR-based capillary sequencing and Sequenom primer-extension mass spectrometry: the existence of the novel allele was confirmed for 270 of the 275 loci. The genotype concordance rate with Sequenom was 99.9% and with capillary sequencing it was 98.6%, excluding heterozygotes (Supplementary Tables 3 and 4). In the case of heterozygous genotypes, deep sequencing gives the allelic ratio, whereas most other *P. falciparum* SNP typing methods give the majority allele or return a missing genotype. The observation of heterozygosity by deep sequencing was correlated with Sequenom's failing to call a majority allele, but when Sequenom made a majority allele call it agreed with deep sequencing data in 94.8% of cases (Supplementary Fig. 5). Capillary sequencing data do not allow allelic ratios to be quantified precisely, but visual inspection of capillary sequence traces was consistent with heterozygous genotype calls in the deep sequencing data (Supplementary Fig. 6). In a separate study to be reported elsewhere, we sequenced 90 laboratory-adapted parasite clones derived from three genetic crosses of *P. falciparum* and determined that the rate of Mendelian errors in the 86k SNP set was 0.05%.

Population genetic analyses were conducted with the 86k SNP set typed in 227 samples as described above. The allele frequency spectrum was dominated by low-frequency variants (Fig. 1 and Supplementary Fig. 7) even when synonymous sites alone were considered, which is consistent with recent population expansion (Supplementary Table 5)[10]. Samples from Africa had a greater number of low-frequency variants than samples from Southeast Asia or Papua New Guinea with or without correction for sample size. Multiple lines of evidence indicate that *P. falciparum* originated in Africa, and loss of low-frequency variation might have occurred as a result of population bottlenecks during migration out of Africa, as in human populations[10,11].

The most likely ancestral state of each SNP was determined from the *P. reichenowi* genome sequence but is difficult to estimate with confidence, because *P. reichenowi* might have diverged from *P. falciparum* relatively recently, and its genome sequence has been determined for only one individual (refs 6, 12 and T. D. Otto, unpublished obervations). There seem to be more SNPs with low-frequency derived (non-ancestral) alleles in Africa than in Southeast Asia or Papua New Guinea (Supplementary Figs 8 and 9). Focusing on SNPs that are private to one continent, those with high derived allele frequency show

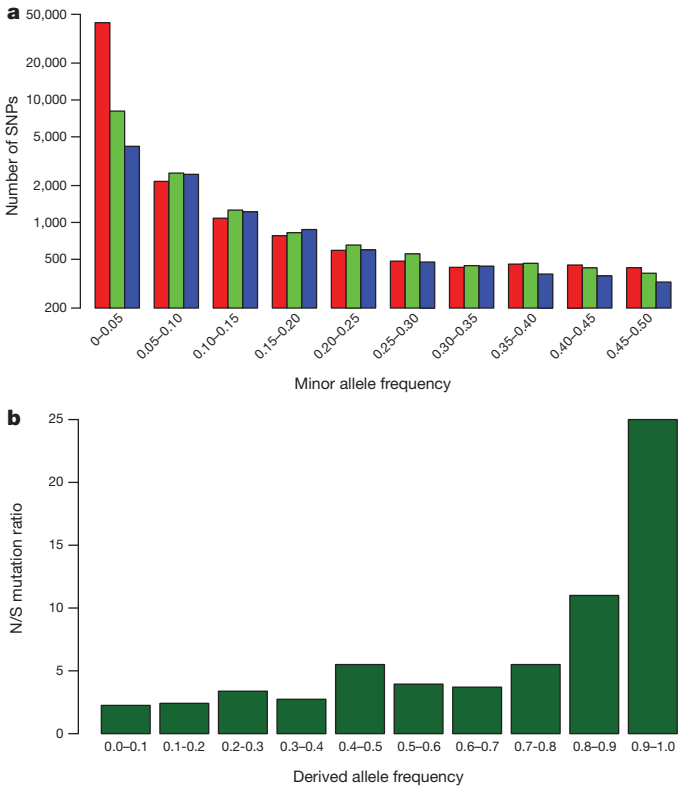


Figure 1 | Allele frequency spectrum of SNPs genotyped in this study. **a**, Minor-allele frequency distribution of 86k SNPs set in samples from different continents: Africa (red), Southeast Asia (green) and Papua New Guinea (blue). The *y* axis shows the number of SNPs in each category of allele frequency. Supplementary Figure 7 shows the data corrected for sample size. **b**, Ratio of non-synonymous (N) to synonymous (S) substitutions, as a function of derived allele frequency for SNPs that are private to either Africa, Southeast Asia or Papua New Guinea.

a considerable excess of non-synonymous substitutions, suggesting that these are largely the result of directional selection (Fig. 1b and Supplementary Fig. 10).

Many SNPs (64%) were observed in only one continent, but most were low-frequency variants and larger sample sizes are needed to determine how many of these are truly private. Corrected for sample size, the number of private SNPs was greatest in East Africa and least in Southeast Asia, both of which comprised cultured samples (Supplementary Fig. 11). Intermediate numbers were observed in West Africa and Papua New Guinea, both of which comprised direct samples. Thus the effect of culturing on SNP ascertainment seems to be relatively small in comparison with the effect of geographical location.

The global population structure of *P. falciparum* shows a clear division by continent (Fig. 2a). Mean fixation index (F_{st}) values between continents ranged from 0.19 to 0.28 (Supplementary Table 6). Population structure within continents is evident from F_{st} values, principal-components analysis (Supplementary Fig. 12) and a neighbour-joining tree (Fig. 2b). All of these methods show greater degree of population structure in Southeast Asia than in West Africa; that is, samples from Cambodia and Thailand form separate clusters, whereas samples from Mali and Burkina Faso are intermixed. These data are consistent with previous evidence that parasite population structure tends to be increased in regions of low or patchy malaria transmission[13].

To understand the hierarchical population structure of *P. falciparum*, methods are needed to quantify the genetic diversity of individual infections relative to the genetic diversity of the parasite population as a whole. With deep sequencing data, we can estimate

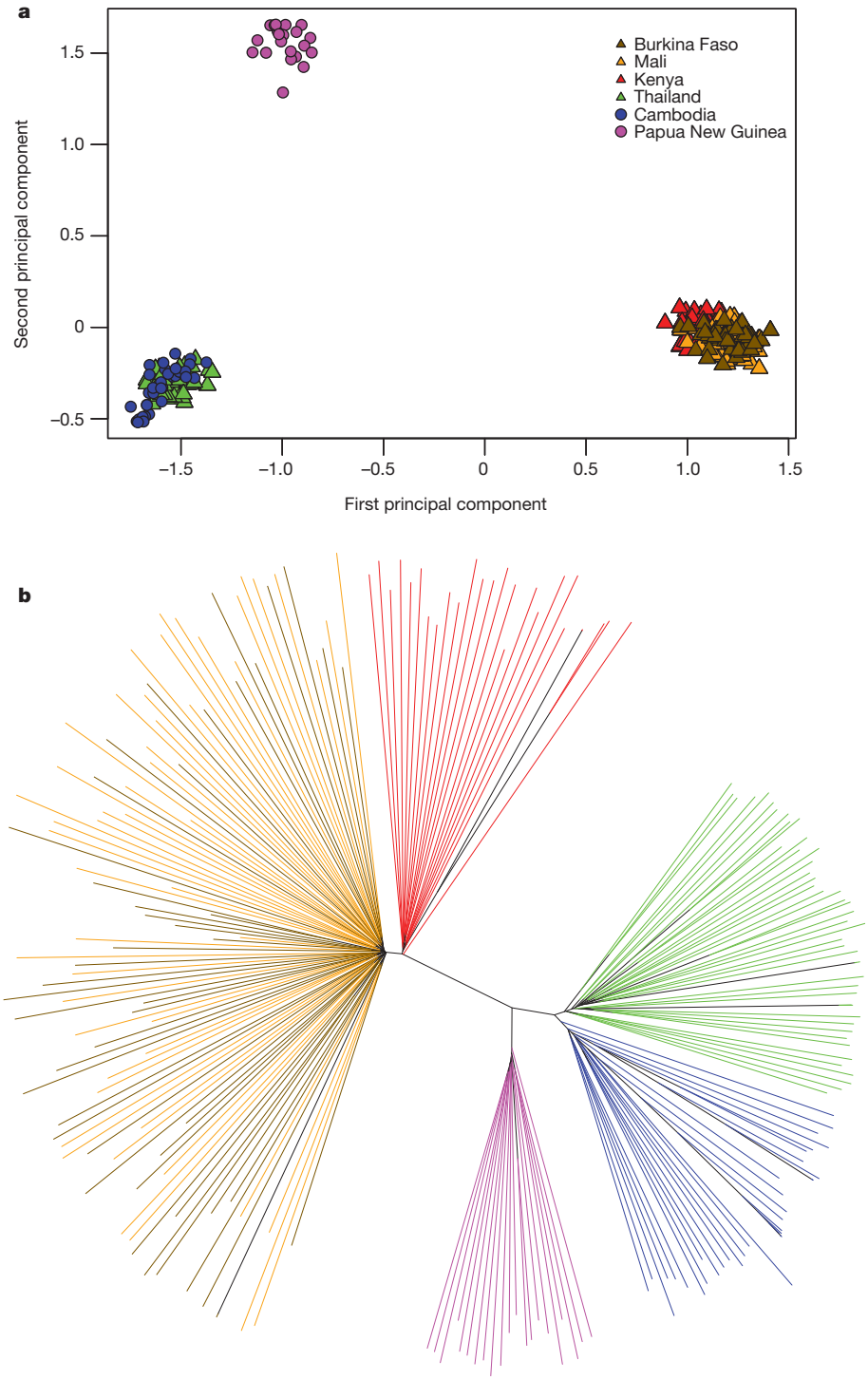


Figure 2 | Representations of a pairwise distance matrix between the 227 samples analysed. **a**, Principal-components analysis. **b**, Unrooted neighbour-joining tree. Leaf branches are coloured (as in **a**) according to the country of origin of the sample.

levels of heterozygosity both within an individual sample (H_w) and within the local parasite population (H_s). For a biallelic SNP, we define H_w as $2p_wq_w$, where p_w and q_w denote the proportions of the two alleles in the sequence reads of an individual sample, and H_s as $2p_sq_s$, where p_s and q_s denote the corresponding population allele frequencies at that geographical location. We observe a strong linear relationship between H_w and H_s when data for all 86k SNPs are aggregated for an individual sample (Fig. 3a and Supplementary Fig. 13). More specifically, each sample shows a linear relationship between H_w and H_s but the gradient of the line varies considerably between samples. This gradient is essentially a genome-wide estimate of H_w/H_s for the sample in question. Thus for each sample we can derive the metric F_{ws}, where

$$F_{ws} = 1 - H_w/H_s$$

This is closely related to Wright's inbreeding coefficient F_{is}, which can be formulated as

$$F_{is} = 1 - H_i/H_s$$

where H_i is the heterozygosity of the individual and H_s is that of the local population[14]. Estimation of F_{is} is of practical relevance for malaria

control, because high rates of inbreeding are thought to favour the emergence of multigenic drug resistance[15,16]. F_{is} is conventionally measured at the oocyst stage of infection—that is, after the parasites have undergone sexual mating within the mosquito and before they develop into separate haploid forms—but this is technically demanding and difficult to implement on a large scale[15,17]. Because parasites undergo sexual mating shortly after the mosquito has ingested blood from an infected person, the level of within-host diversity determines the potential for inbreeding or outcrossing in the next generation. Thus F_{ws} values observed in blood samples provide a proxy indicator of inbreeding rates in the population. The precise relationship to inbreeding rates quantified in oocysts merits further investigation. We report elsewhere a study of how F_{ws} relates to standard methods of estimating multiplicity of infection[18].

We observe marked differences in F_{ws} between locations (Fig. 3b). High levels of F_{ws} (0.95 or more) were much more common in Papua New Guinea (89% of samples) than in West Africa (38%), with intermediate rates in Southeast Asia (67%) and East Africa (63%). Culturing might affect F_{ws} estimation, but the samples from Papua New Guinea and West Africa were not cultured. In general, high levels of inbreeding tend to be associated with low transmission intensity[13], and these data are therefore somewhat surprising because the entomological inoculation rate has been estimated to lie in the range 45–293 in Madang in Papua New Guinea[19], where the Papua New Guinea samples were collected, in contrast with 140–389 in Burkina Faso[19], about 6 in rural areas of Cambodia[20] and about 1 on the Thailand–Burma border[21]. Although the entomological inoculation rate can be highly variable within a locality and these estimates are indicative, it seems unlikely that the high levels of F_{ws} in Papua New Guinea are primarily due to low transmission intensity. An alternative explanation is that, in this geographical region, people tend to live in small isolated communities, which might reduce the likelihood of infection with parasites of different genetic types. The small size of the Papua New Guinea sample provides limited information about local parasite population structure (Supplementary Fig. 14), but previous studies indicate that this is very high in some villages within this area of Papua New Guinea[22].

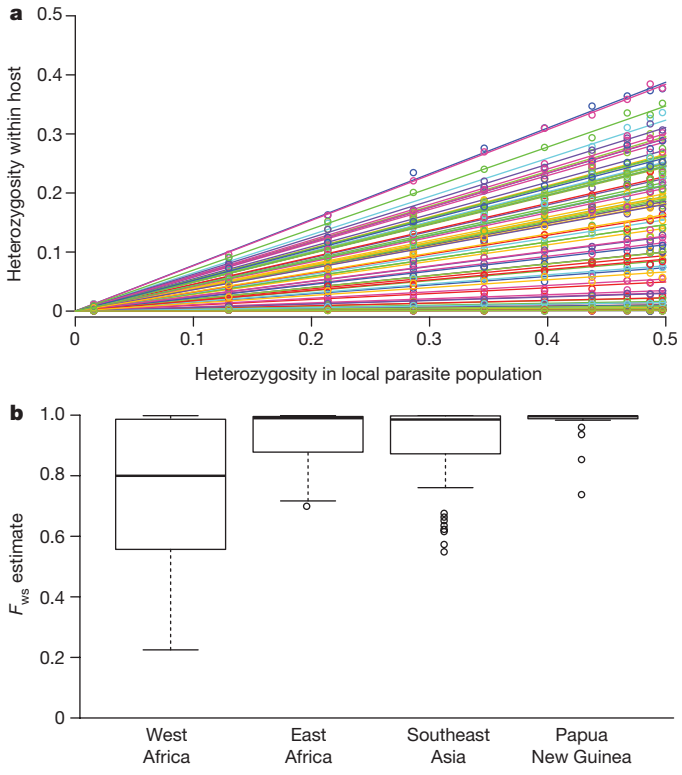


Figure 3 | Quantification of within-host diversity. a, Relationship between within-host heterozygosity (H_w) and heterozygosity in the local parasite population (H_s) for all samples in the West African population. Each line represents a different sample, whose within-host heterozygosity values were averages across all SNPs, categorized according to their heterozygosity in the local parasite population. Separate plots for each population are shown in Supplementary Fig. 13. **b**, Box plot showing the distribution of F_{ws} estimates in samples from each of the four populations.

These data allow linkage disequilibrium in the *P. falciparum* genome to be estimated with greater precision than has previously been possible. In particular, we can begin to distinguish linkage disequilibrium due to haplotype structure, which decays with distance in the genome, from linkage disequilibrium due to population structure, which is independent of distance in the genome (see Supplementary Methods, Supplementary Tables 7 and 8 and Supplementary Figs 15–17). Averaged across the genome, after correcting for population structure and other confounders, we find that r^2 decays to less than 0.1 within 1 kilobase (kb) in all populations studied here, whereas D' decays to less than 0.1 within about 1 kb in West Africa and East Africa, and within 50 kb in Southeast Asia and Papua New Guinea (Supplementary Fig. 18). These findings imply that high levels of haplotypic diversity exist at all of these locations, despite low transmission intensity and high rates of inbreeding at some locations. This might be partly due to the high rate of meiotic recombination in *P. falciparum*, previously estimated to be about 17 kb per centimorgan[23]. It is also possible that much of the haplotypic diversity seen in contemporary *P. falciparum* populations has ancient origins, and arose in Africa before *P. falciparum* was spread around the world by human migration. This would be analogous to the situation that is seen in human populations, in which migration out of Africa was associated with a series of population bottlenecks, which have led to a reduction in haplotypic diversity in descendant populations around the world[11]. The higher levels of linkage disequilibrium observed in Southeast Asia and Papua New Guinea than in West Africa and East Africa are consistent with both of these possibilities.

A web application is provided for browsing, querying and downloading information about all of the SNPs genotyped in this study and their allele frequencies in different geographical regions (http://www.malariagen.net/resource/10). It can be used, for example, to view regional patterns of variation in known antimalarial drug resistance genes: from these data it is immediately apparent that the *pfcrt* K76T allele has markedly different haplotypic backgrounds in Southeast Asia and in Papua New Guinea, consistent with previous evidence that chloroquine resistance has evolved independently in multiple locations (Supplementary Table 9)[1,24]. It can also be used to search for genes that are highly differentiated between geographical regions (Supplementary Tables 10 and 11). For example, two genes that affect the fertility of gametocytes, *Pfs230* and *Pf47*, are among the most highly differentiated loci in this data set[25]. Two SNPs in *Pfs230* codon 1566 result in three amino-acid variants: N (widespread), T (private to Southeast Asia, frequency 0.87) and K (private to Africa, frequency 0.79). Codon variant T236I of *Pf47* has a fixed difference between Africa and other populations. These data lend weight to previous reports of extreme differentiation in *Pf47* and the related gene *Pfs48/45* (ref. 26), which is suggested to be due to evolutionary selection of gamete recognition and compatibility. Another example is codon variant F368S of the putative transporter gene *PFA0245w* (ref. 27), which has a fixed difference between Papua New Guinea and other populations, raising the question of whether this has a function in drug resistance; it is also noteworthy that the *Plasmodium berghei* orthologue of this gene is critical for sexual development of the parasite[28].

These data are the first stage in the development of methods for population-based genome sequencing of *P. falciparum*. Work is ongoing to increase the number of SNPs that can be reliably genotyped, and to develop accurate methods for typing indels, copy number polymorphisms and large structural variations. Future studies will benefit from new methods to reduce the effects of AT bias on sequencing library preparation[29,30], and the increasing length and accuracy of sequencing

reads will allow greater access to highly polymorphic regions of the genome. Such technical advances will enable an expanding range of applications, for example high-resolution analyses of local population structure to explore models of space–time clustering and immunological strain selection.

Genome sequencing of parasites in clinical blood samples is an important step towards translation to public health applications, for example developing effective genetic markers to track the spread of antimalarial drug resistance and to monitor evolutionary changes in the parasite population[7,8]. There is a need to develop protocols, tools and resources and to enable researchers in malaria endemic countries to integrate parasite genome sequencing into clinical and epidemiological investigations, and to facilitate open-access sharing of large-scale population genomic data.

METHODS SUMMARY

Blood samples from malaria patients were collected with informed consent after approval by local ethics committees. Parasite DNA was extracted from blood samples after leukocyte depletion to minimize contamination with human DNA, or after short-term culture *in vitro*. Samples with less than 60% human DNA contamination were sequenced with an Illumina Genome Analyser. Sequence reads of length 37–76 base pairs were aligned to the 3D7 reference sequence[3] using the bwa and samtools algorithms, and then with the more stringent SNP-o-matic algorithm that allowed for SNPs discovered in the first step. This gave 868,117 potential SNPs, including 74% (71,608/96,527) of SNPs previously identified in the PlasmoDB 5.5 database.

Various quality-control steps were applied. We discarded potential SNPs with insufficient evidence, those in non-coding regions, and those in coding regions with sequencing coverage outside the 15th centile and the 85th centile of read depth. To minimize alignment errors, we scored each position in the reference genome for its degree of uniqueness, and excluded positions that were liable to give false heterozygous genotypes. We analysed levels of heterozygosity across all samples, discarding positions where heterozygosity was inconsistent with population allele frequencies. Genotypes were determined at positions with at least five reads, resulting in a set of 86,158 biallelic SNPs that could be genotyped with low missingness in 227 samples.

Five methods were used to validate genotyping calls: Sequenom primer-extension mass spectrometry, PCR-based capillary sequencing, Illumina GoldenGate array, high-density NimbleGen microarray, and analysis of error rates in genotypes from *P. falciparum* genetic crosses. Allele frequencies were determined in four populations, deriving ancestral alleles from comparison with *P. reichenowi* sequences wherever possible. SNPs were classified in accordance with PlasmoDB 5.5 functional annotations. Principal-components analysis and phylogeny analysis were performed using R language libraries, and custom R and Java programs were used for other data analysis.

Received 24 December 2010; accepted 30 April 2012.
Published online 13 June 2012.

1. Wootton, J. C. *et al.* Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*. *Nature* **418,** 320–323 (2002).
2. Dondorp, A. M. *et al.* Artemisinin resistance in *Plasmodium falciparum* malaria. *N. Engl. J. Med.* **361,** 455–467 (2009).
3. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).
4. Mu, J. *et al.* Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. *Nature Genet.* **39,** 126–130 (2007).
5. Volkman, S. K. *et al.* A genome-wide map of diversity in *Plasmodium falciparum*. *Nature Genet.* **39,** 113–119 (2007).
6. Jeffares, D. C. *et al.* Genome variation and evolution of the malaria parasite *Plasmodium falciparum*. *Nature Genet.* **39,** 120–125 (2007).
7. Neafsey, D. E. *et al.* Genome-wide SNP genotyping highlights the role of natural selection in *Plasmodium falciparum* population divergence. *Genome Biol.* **9,** R171 (2008).
8. Mu, J. *et al. Plasmodium falciparum* genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nature Genet.* **42,** 268–271 (2010).
9. Auburn, S. *et al.* An effective method to purify *Plasmodium falciparum* DNA directly from clinical blood samples for whole genome high-throughput sequencing. *PLoS ONE* **6,** e22213 (2011).
10. Joy, D. A. *et al.* Early origin and recent expansion of *Plasmodium falciparum*. *Science* **300,** 318–321 (2003).
11. Li, J. Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319,** 1100–1104 (2008).
12. Prugnolle, F. *et al.* African great apes are natural hosts of multiple related malaria species, including *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **107,** 1458–1463 (2010).
13. Anderson, T. J. *et al.* Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Mol. Biol. Evol.* **17,** 1467–1482 (2000).
14. Hartl, D. & Clark, A. G. *Principles of population genetics* 4th edn (Sinauer, 2007).
15. Paul, R. E. *et al.* Mating patterns in malaria parasite populations of Papua New Guinea. *Science* **269,** 1709–1711 (1995).
16. Dye, C. & Williams, B. G. Multigenic drug resistance among inbred malaria parasites. *Proc. R. Soc. Lond. B* **264,** 61–67 (1997).
17. Hill, W. G., Babiker, H. A., Ranford-Cartwright, L. C. & Walliker, D. Estimation of inbreeding coefficients from genotypic data on multiple alleles, and application to estimation of clonality in malaria parasites. *Genet. Res.* **65,** 53–61 (1995).
18. Auburn, S. *et al.* Characterization of within-host *Plasmodium falciparum* diversity using next-generation sequence data. *PLoS ONE* **7,** e32891 (2012).
19. Smith, D. L., Drakeley, C. J., Chiyaka, C. & Hay, S. I. A quantitative analysis of transmission efficiency versus intensity for malaria. *Nature Commun.* **1,** 108 (2010).
20. Trung, H. D. *et al.* Malaria transmission and major malaria vectors in different geographical areas of Southeast Asia. *Trop. Med. Int. Health* **9,** 230–237 (2004).
21. Paul, R. E. *et al.* Genetic analysis of *Plasmodium falciparum* infections on the north-western border of Thailand. *Trans. R. Soc. Trop. Med. Hyg.* **93,** 587–593 (1999).
22. Schultz, L. *et al.* Multilocus haplotypes reveal variable levels of diversity and population structure of *Plasmodium falciparum* in Papua New Guinea, a region of intense perennial transmission. *Malar. J.* **9,** 336 (2010).
23. Su, X. *et al.* A genetic map and recombination parameters of the human malaria parasite *Plasmodium falciparum*. *Science* **286,** 1351–1353 (1999).
24. Mehlotra, R. K. *et al.* Evolution of a unique *Plasmodium falciparum* chloroquine-resistance phenotype in association with *pfcrt* polymorphism in Papua New Guinea and South America. *Proc. Natl Acad. Sci. USA* **98,** 12689–12694 (2001).
25. van Dijk, M. R. *et al.* Three members of the 6-cys protein family of *Plasmodium* play a role in gamete fertility. *PLoS Pathog.* **6,** e1000853 (2010).
26. Anthony, T. G., Polley, S. D., Vogler, A. P. & Conway, D. J. Evidence of non-neutral polymorphism in *Plasmodium falciparum* gamete surface protein genes *Pfs47* and *Pfs48/45*. *Mol. Biochem. Parasitol.* **156,** 117–123 (2007).
27. Martin, R. E., Henry, R. I., Abbey, J. L., Clements, J. D. & Kirk, K. The 'permeome' of the malaria parasite: an overview of the membrane transport proteins of *Plasmodium falciparum*. *Genome Biol.* **6,** R26 (2005).
28. Boisson, B. *et al.* The novel putative transporter NPT1 plays a critical role in early stages of *Plasmodium berghei* sexual development. *Mol. Microbiol.* **81,** 1343–1357 (2011).
29. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods* **6,** 291–295 (2009).
30. Oyola, S. O. *et al.* Optimizing Illumina Next-Generation Sequencing library preparation for extremely AT-biased genomes. *BMC Genomics* **13,** 1 (2012).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank G. Dougan and N. Day for support, and T. Anderson and M. Mackinnon for comments. The sequencing and analysis components of this study were supported by the Wellcome Trust through Sanger Institute core funding (077012/Z/05/Z; 098051) and a Strategic Award (090770/Z/09/Z); the Medical Research Council (MRC) through the MRC Centre for Genomics and Global Health (G0600718) and an MRC Professorship to D.P.K. (G19/9). Other parts of this study were partly supported by the Wellcome Trust including core support to the Wellcome Trust Centre for Human Genetics (075491/Z/04; 090532/Z/09/Z); the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health; and a Howard Hughes Medical Institute International Scholarship (55005502) to A.D.

Author Contributions S.A., S.C., A.D., O.D., I.Z., J.-B.O., P.M., I.M., P.S., A.N., S.B., S.M.K., K.M., H.J., X.-Z.S., C.A., R.F., D.S., F.N., M.I., N.J.W., L.A.-E., C.S., V.M., D.M., A.A.-N. and D.J.C. performed field and laboratory studies to obtain *P. falciparum* samples for sequencing. S.A., S.C., M.S., E.A., D.A., E.D., S.O., M.A.Q., D.J.T., B.M., C.I.N. and M.B. developed and implemented methods for sample processing and sequencing library preparation. J.A.-G., M.M., O.M., G.M., V.R.R. and D.J. developed software for data management and visualization. K.A.R., C.H., A.J., K.R., J.C.T., M.T.F., S.C., S.A., D.A., C.I.N. and M.B. performed validation experiments. C.V.P., S.T.-H. and C.R. contributed to development of the project. B.M., M.B., C.I.N. and J.C.R. provided project management and oversight. O.M., M.M., D.P.K., J.O.'B. and T.G.C. conducted data analyses. D.P.K. and O.M. developed the F_{ws} metric. D.P.K., O.M. and M.M. wrote the manuscript and collated comments from all authors. S.A. and S.C. made equal contributions.

Author Information All sequence data are available online at the European Nucleotide Archive (ENA); accession numbers are listed in Supplementary Table 12. An online catalogue of SNPs and allele frequencies is available at http://www.malariagen.net/resource/10. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to D.P.K. (dominic@sanger.ac.uk).

LETTER

doi:10.1038/nature11207

SHARP1 suppresses breast cancer metastasis by promoting degradation of hypoxia–inducible factors

Marco Montagner[1], Elena Enzo[1], Mattia Forcato[2], Francesca Zanconato[1], Anna Parenti[3], Elena Rampazzo[4], Giuseppe Basso[4], Genesio Leo[5], Antonio Rosato[6], Silvio Bicciato[2], Michelangelo Cordenonsi[1] & Stefano Piccolo[1]

The molecular determinants of malignant cell behaviours in breast cancer remain only partially understood[1]. Here we show that SHARP1 (also known as BHLHE41 or DEC2) is a crucial regulator of the invasive and metastatic phenotype in triple-negative breast cancer (TNBC), one of the most aggressive types of breast cancer. SHARP1 is regulated by the p63 metastasis suppressor and inhibits TNBC aggressiveness through inhibition of hypoxia-inducible factor 1α (HIF-1α) and HIF-2α (HIFs). SHARP1 opposes HIF-dependent TNBC cell migration *in vitro*, and invasive or metastatic behaviours *in vivo*. SHARP1 is required, and sufficient, to limit expression of HIF-target genes. In primary TNBC, endogenous SHARP1 levels are inversely correlated with those of HIF targets. Mechanistically, SHARP1 binds to HIFs and promotes HIF proteasomal degradation by serving as the HIF-presenting factor to the proteasome. This process is independent of pVHL (von Hippel–Lindau tumour suppressor), hypoxia and the ubiquitination machinery. SHARP1 therefore determines the intrinsic instability of HIF proteins to act in parallel to, and cooperate with, oxygen levels. This work sheds light on the mechanisms and pathways by which TNBC acquires invasiveness and metastatic propensity.

Breast cancer is a heterogeneous disease, both biologically and clinically[2]. An important parameter in the clinical management of breast cancer patients is the expression of steroid hormone receptors (oestrogen and progesterone receptors) and ERBB2 (also known as HER2) overexpression and/or amplification; effective tailored therapies have in fact been developed for patients with hormone receptor-positive or HER2-positive diseases[2]. In contrast, TNBC is defined merely by the lack of expression of oestrogen receptor, progesterone receptor and HER2, and thus the category includes tumours that are clinically and pathologically diverse, for which we strive to identify tumour-addicted molecular pathways[1]. Understanding TNBC is of pivotal importance not only in light of the current lack of therapeutic options but also because TNBC accounts for some of the most aggressive types of breast cancers, marked by high rates of relapse, visceral metastases and early death[1].

TAp63 (one of the two main isoforms encoded by *p63*) has recently emerged as a key suppressor of invasive and metastatic cell behaviours in breast cancer, and other tumour types[3–5]; for example, mice that have TAp63 genetically ablated develop aggressive and metastatic carcinomas[4]. We recently identified two genes, *SHARP1* and *CCNG2* (cyclin G2), that are downstream of TAp63α (a sub-isoform of TAp63) in breast cancer cells[3]. To investigate the involvement of this pathway in promoting malignancy and metastatic spread in TNBC, we started by collecting a cohort of 250 primary TNBC samples from eight clinically annotated gene-expression data sets (see Methods and Supplementary Tables 1, 2 and 3), and examined whether the expression of *SHARP1* and *CCNG2* has prognostic value in TNBC. For this, we defined two groups of TNBC with high and low levels of *SHARP1* and *CCNG2* expression, respectively (Supplementary Fig. 1). Crucially, using univariate Kaplan–Meier analyses, individuals in the TNBC group with low *SHARP1* and *CCNG2* expression were shown to have a significantly higher probability of developing metastasis and of reduced survival (Fig. 1a and Supplementary Fig. 1). Moreover, we discriminated between different p63 carboxy-terminal variants using the probe sets present on the TNBC microarrays, and found that the levels of *SHARP1* and *CCNG2* expression correlated with the levels of p63α/β, but not of p63γ (Supplementary Fig. 2).

The fact that expression of just two genes could be prognostic in TNBC indicates that in addition to their utility as markers they may have functional roles in tumour aggressiveness. To gain insights into the mechanisms by which SHARP1 and CCNG2 are linked to malignant progression, we investigated whether their expression could be linked to other known tumorigenic pathways by gene set enrichment analyses (GSEA, see Methods). Specifically, we searched in tumour samples for statistical associations between low *SHARP1* and *CCNG2* expression, and other gene 'signatures' that register elevated activity of various signalling pathways or dysregulated cellular processes, for a total of 254 signatures (Supplementary Tables 5 and 6). Interestingly, signatures of TGFβ activity and p53 mutation were most strongly associated with low *SHARP1* and *CCNG2* expression (Fig. 1b). Hence, clinical data confirm that mutant p53 and TGFβ are both involved in the regulation of *SHARP1* and *CCNG2*, as previously shown in the well-established TNBC cellular model MDA-MB-231 (also known as MDA-231) (ref. 3). We then focused on the signature with the second strongest association with low *SHARP1* and *CCNG2* expression: a signature that denotes high HIF activity (Fig. 1b). HIFs have been involved in several aspects of malignancy in several tumour types[6–8]; in breast cancer, a large body of clinical data shows that high levels of HIF-1α or HIF-2α in tissue samples are linked to poor prognosis and high patient mortality[8,9]. Several results strongly support the role of HIFs in TNBC: first, a signature of HIF activity predicted metastasis proclivity in our cohort of TNBC (Supplementary Fig. 3); second, combined loss of HIF-1α and HIF-2α severely impaired lung colonization of MDA-231 cells injected in the tail vein of immunocompromised mice (Fig. 2a; refs 10, 11); and third, short interfering RNA (siRNA)-mediated depletion of HIF-1α or HIF-2α potently inhibited trans-well migration that is triggered by TGFβ (Supplementary Fig. 4).

We next examined whether SHARP1 or CCNG2 are causal for the repression of HIFs. Interestingly, a physical association between overexpressed HIF-1α and SHARP1 has been previously noted in transfected COS7 cells[12]. We found a robust physical association between HIF-1α and SHARP1 at endogenous levels in independent TNBC cell lines, such as MDA-231, Hs578T and SUM159 (Fig. 1c, data not shown). Conversely, no interaction was detected between HIF-1α and CCNG2 (data not shown); moreover, in contrast with SHARP1 (see below), CCNG2 overexpression had no inhibitory effect on the

[1]Department of Medical Biotechnologies, University of Padua School of Medicine, Viale Colombo 3, 35131 Padua, Italy. [2]Center for Genome Research, Department of Biomedical Sciences, University of Modena and Reggio Emilia, Via G. Campi 287, 41100 Modena, Italy. [3]Department of Medical Diagnostic Science and Special Therapies, Section of Pathology, University of Padua, Viale Gabelli 2, 35126 Padua, Italy. [4]Clinical and Experimental Hematology, Department of Pediatrics, University of Padova, Via Giustiniani 3, 35128 Padova, Italy. [5]Division of Anatomic Pathology, Hospital San Bassiano, Via dei Lotti 40, 36061 Bassano del Grappa, Italy. [6]Department of Surgery, Oncology and Gastroenterology, and Istituto Oncologico Veneto IRCCS, Via Gattamelata 64, 35126 Padua, Italy.

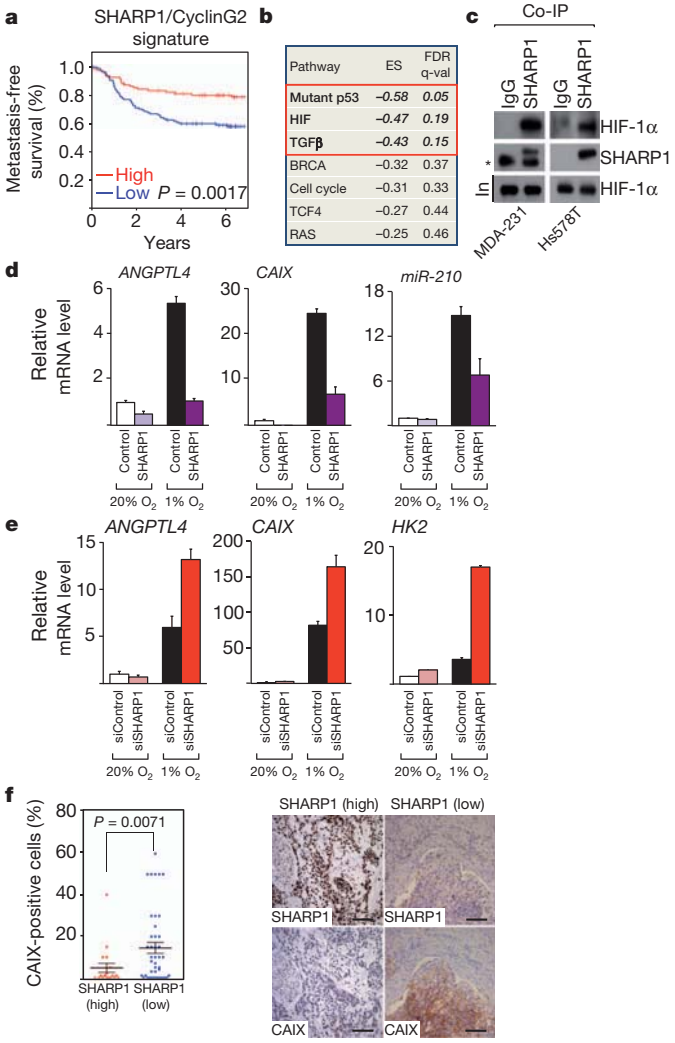


Figure 1 | SHARP1 is an inhibitor of HIF activity. **a**, Kaplan–Meier graph representing the probability of metastasis-free survival in TNBC patients, stratified according to high or low expression levels of *SHARP1* and *CCNG2* (*SHARP1*/*CCNG2* signature). In multivariate analysis, the *SHARP1*/*CCNG2* signature is found to be an independent predictor of survival that adds new prognostic information to established clinical predictors such as size and age (Supplementary Table 4). **b**, Gene set enrichment analysis (GSEA) for association between high or low *SHARP1* and *CCNG2* expression values and gene sets denoting the activation of specific signalling pathways. The enrichment of HIF-target genes in tumours with low *SHARP1* and *CCNG2* expression was confirmed using an independent statistical method (Supplementary Table 7). Bold, signatures that reach statistical significance. ES, enrichment score; FDR, false discovery rate. **c**, Co-immunoprecipitation (Co-IP) of endogenous SHARP1 with endogenous HIF-1α, from extracts of MDA-231 and Hs578T cells. Asterisk, a background band resulting from the cross-reaction of immunoglobulins (IgGs). In, input. **d**, Quantitative polymerase chain reaction (qPCR) analyses of selected HIF targets in MDA-231 cells stably overexpressing empty vector (Control) or SHARP1, and incubated in low (1%) or normal (20%) oxygen levels. Expression levels are relative to glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*), data are normalized to lane 1 (for analyses of additional HIF targets, see Supplementary Fig. 6; and for control of SHARP1 overexpression, see Fig. 3a). The mean and s.d. of one representative experiment, out of three independent experiments performed in triplicate, are shown. **e**, qPCR analyses of selected HIF targets in MII cells transfected with the indicated siRNAs (Supplementary Table 10), and incubated at the indicated oxygen levels (for analyses of additional HIF targets, see Supplementary Fig. 7). The mean and s.d. of one representative experiment, out of three independent experiments performed in triplicate, are shown. **f**, Immunohistochemistry was used to stain TNBC samples for SHARP1 and the HIF target CAIX ($n = 62$). Left, error bars represent mean and s.e.m. The *P* value was calculated using the Student's *t*-test. Right, representative immunohistochemistry. Scale bars, 100 μm.

activation of HIF targets (data not shown), prompting us to focus on SHARP1 for further analyses.

SHARP1 levels seem elevated in the TNBC non-metastatic MCF10Atk1 cells (also known as MII cells) compared to more aggressive MDA-231cells (Supplementary Fig. 5). Taking advantage of this differential SHARP1 expression, we reasoned that if SHARP1 opposes HIF activity, gain of SHARP1 in MDA-231 cells should blunt HIF transcriptional responses, and conversely, loss of SHARP1 in MII cells should enhance HIF-dependent responses. HIF-1α and HIF-2α regulation occurs in response to microenvironmental oxygen levels[7]. Under hypoxia, HIF levels are stabilized to activate a key set of target genes, such as angiopoietin-like 4 (*ANGPTL4*), *LOXL2* and *miR-210,* that have been involved in multiple steps of the metastatic cascade of oestrogen -negative breast cancer cells[13–15]. As expected, these genes were upregulated by hypoxia in control MDA-231 cells, but gain of SHARP1 dampened these inductions (Fig. 1d and Supplementary Fig. 6). Similarly, inductions of HIF targets related to tumour metabolism, such as carbonic anhydrase IX (*CAIX*; also known as *CA9*), hexokinase 2 (*HK2*) and pyruvate dehydrogenase kinase 1 (*PDK1*), or the autophagy and stress regulators *BNIP3* and *NDRG1* (ref. 7), were also inhibited by SHARP1 (Fig. 1d and Supplementary Fig. 6). Thus, sustaining SHARP1 expression is sufficient to oppose HIF responses. Conversely, in MII cells, loss of SHARP1 strongly cooperated with a pulse of hypoxia to upregulate the HIF targets vascular endothelial growth factor A (*VEGFA*), *CAIX*, *ANGPTL4*, *HK2* and *PDK1* (Fig. 1e and Supplementary Fig. 7). This indicates that SHARP1 is required to limit HIF activity and is an endogenous buffer against the effects of hypoxia.

To investigate in an unbiased way and at a genome-wide level whether SHARP1 and HIFs control a significantly overlapping set of genes, transcriptomic profiles were obtained from cells stably expressing either short hairpin green fluorescent protein (shGFP) or shRNAs against HIF-1α and HIF-2α, and from cells overexpressing SHARP1. We identified two independent lists of genes differentially expressed after SHARP1 overexpression or after depletion of HIFs (Supplementary Tables 8 and 9, and Supplementary Fig. 8). When these lists were compared, we found a highly statistically significant overlap between the two lists (Fischer's test, $P < 10^{-73}$), supporting the idea that SHARP1 is a global inhibitor of HIF-1α and HIF-2α gene responses. From these microarrays, we generated a list of genes repressed by SHARP1 (Supplementary Table 6) and found that this signature has prognostic value in TNBC data sets. Tumours expressing high levels of such genes (that is, low SHARP1 activity signature) (Supplementary Fig. 9) display increased propensity to distant metastasis than tumours expressing low levels of the same signature (that is, those retaining high SHARP1 activity). Cox multivariate analyses revealed that the signature of SHARP1-repressed genes did not add prognostic information when combined to a signature of high HIFs activity ($P = 0.4434$); a similar result was obtained from the combination of the *SHARP1* and *CCNG2* signature with HIF activity signature ($P = 0.3072$), indicating that the prognostic value of SHARP1 activity is contained in the prognostic value of HIFs.

These results prompted us to verify the inverse correlation between HIF activity and *SHARP1* expression in primary human TNBC. Levels of *VEGF* transcripts and of other HIF targets were found to decline in tumours displaying increasing levels of *SHARP1* (Supplementary Fig. 10). Notably, immunohistochemical analysis of an independent cohort of 62 TNBC diagnosed at our institution showed an inverse relationship between SHARP1 and CAIX, a validated immunohistological marker of HIF activity in invasive breast cancer[16] (Fig. 1f).

We then investigated the functional relevance of the SHARP1–HIF axis for malignant behaviour of TNBC cells. *In vivo*, both the depletion of HIFs and gain of SHARP1 opposed lung colonization of MDA-231 cells after tail-vein injection in recipient mice (Fig. 2a). SHARP1-mediated inhibition was partially rescued by overexpression of PA-HIF-1α, a constitutively active and pVHL-insensitive HIF-1α[7]

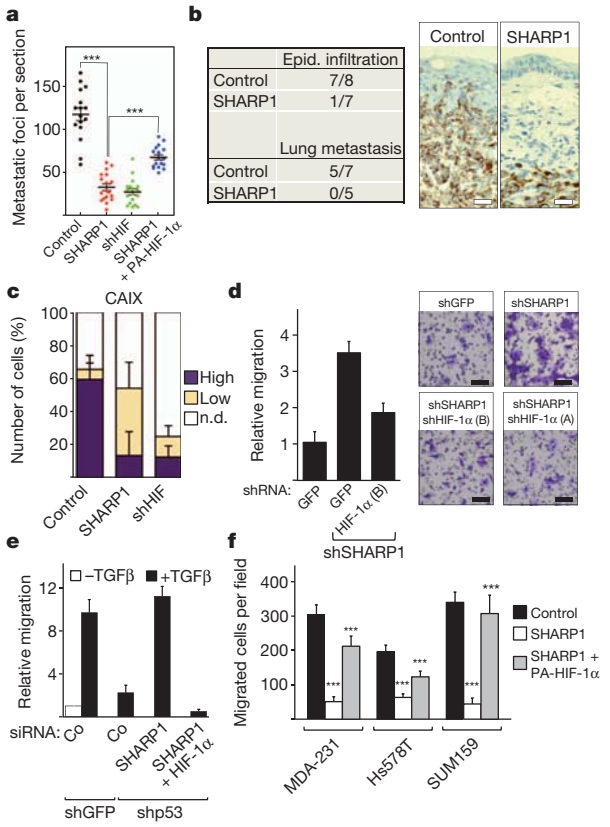


Figure 2 | SHARP1 is a suppressor of invasion, migration and metastasis by inhibiting HIFs. **a**, Lung colonization assay of mice intravenously injected with MDA-231 cells stably transduced with the indicated expression vectors or shHIFs. The control is an empty vector (pLPCX). The plot shows the number of metastatic foci per section, error bars represent mean and s.e.m. (two sections per lung, $n > 8$ mice; $***P < 0.0001$, based on Student's t-test). **b**, Quantification of lung colonization and local invasiveness of primary tumours emerging from orthotopically injected control and SHARP1-overexpressing MDA-231 cells (left panel), and representative images (right panel). Tumours were stained for human cytokeratin. Scale bars, 50 μm. Epid, epidermal. **c**, The HIF-target CAIX protein level was analysed by fluorescent immunohistochemistry in tumours ($n = 7$) emerging after fat pad injection of Control and SHARP1-expressing MDA-231 in mice. Depletion of HIF-1α and HIF-2α served as a positive control for CAIX inhibition (see Supplementary Fig. 11 for comparable results with Glut1 and representative images). n.d., signal not detectable. The error bars represent mean and s.e.m. **d**, Trans-well migration assay of MII cells, a non-metastatic TNBC model system, stably expressing the indicated shRNAs (see Supplementary Table 10). The plot shows the quantification of the area covered by the migrated cells, relative to the first lane that was set to 1. A similar result was obtained with an independent shRNA for HIF-1α (data not shown; see Supplementary Table 10, sequence A). A, B, two sequences targeting HIF-1α (these are given in Supplementary Table 10) Right panels, representative images of the filters (see Supplementary Fig. 14 for knockdown controls). Scale bars, 50 μm. The mean and s.d. of one representative experiment, out of three independent experiments performed in triplicate, are shown. **e**, Trans-well migration assay in MDA-231 cells. Cells expressing shRNA targeting GFP or an shRNA targeting mutant p53 were transiently transfected with the indicated siRNAs (siSHARP1 is sequence B in Supplementary Table 10). TGFβ1 was used to trigger cell migration. Graphs show the quantification of cells that passed through the filter, relative to the first lane that was set to 1. Similar results were obtained by transfecting HIF-1β (Supplementary Fig. 15). The mean and s.d. of one representative experiment, out of five independent experiments performed in triplicate, are shown. **f**, Quantification of wound-healing assays of multiple TNBC cell lines stably transfected with the indicated plasmids (see Supplementary Fig. 16 for representative images). The mean and s.d. of one representative experiment, out of three independent experiments performed in triplicate, are shown ($***P < 0.0001$, Student's t-test; the SHARP1 group was compared to the control group, and the group of SHARP1 with HIF-1α was compared to the SHARP1 group).

(Fig. 2a). When cells were injected orthotopically as a primary tumour in the mammary fat pad, we also noticed a dramatic difference in local invasiveness and distant metastasis between control and SHARP1-expressing tumours (Fig. 2b). In agreement with the role of SHARP1 as a HIF inhibitor, experimental tumours derived from SHARP1-expressing cells showed a reduction of the HIF targets CAIX and Glut1 to levels comparable to tumours emerging from HIF-depleted cells (Fig. 2c and Supplementary Fig. 11). Notably, gain of SHARP1 recapitulates the response of TAp63α, as TAp63α-expressing MDA-231 cells also show a severely impaired metastatic spread in tail-vein assays or after their orthotopic injection into the mammary fat pad (Supplementary Fig. 12). This is consistent with *SHARP1* levels being dependent on *TAp63α* in TNBC cell lines (Supplementary Fig. 13).

The above experiments indicate that by antagonizing HIFs, SHARP1 may control migratory and invasive cell behaviours in tumours. To substantiate further this conclusion, we monitored the relevance of SHARP1 for HIF-dependent cell migration *in vitro*. Loss of SHARP1 leads to increased trans-well migration in MII cells, and this effect was rescued by concomitant depletion of HIF-1α (Fig. 2d). As shown in Fig. 2e, similar results were obtained in MDA-231 cells depleted of mutant p53 (a treatment that relieves the inhibition of endogenous TAp63α activity in these cells, raising SHARP1 levels[3,5,17,18]). To extend the validity of this epistasis, gain of HIFs opposes the anti-migratory effects of gain of SHARP1 in other TNBC cell lines, such as SUM159 and Hs578T. SHARP1-expressing cells were unable to migrate in a wound-healing assay, and this could be rescued by over-expression of PA-HIF-1α or PA-HIF-2α (Fig. 2f and Supplementary Figs 16 and 17). These responses occurred in the absence of any notable effect on cell growth (Supplementary Fig. 18, data not shown).

The data presented so far establish SHARP1 as a physiological inhibitor of HIF function. We next sought to determine the mechanism of this inhibition. HIF regulation occurs mainly at the level of protein stability[7]. Notably, SHARP1 overexpression greatly reduced endogenous HIF-1α protein levels, a result confirmed in three independent TNBC cell lines (MDA-231, SUM159 and Hs578T) (Fig. 3a, b). SHARP1 was equally effective against HIF-2α (Supplementary Fig. 19). Intriguingly, the effect of SHARP1 was independent from oxygen levels, as downregulation of HIF-1α protein occurred highly efficiently in cells cultured in both normoxic and in hypoxic conditions (Fig. 3a). In experimental tumours derived from orthotopically injected MDA-231 cells, SHARP1 also downregulated HIF-1α protein levels to those of cells expressing shHIFs (Fig. 3c and Supplementary Fig. 20).

We then examined whether endogenous SHARP1 is a relevant inhibitor of HIF-1α protein levels. For this, we depleted SHARP1 from MII cells and MDA-231 cells. SHARP1 depletion induced robust HIF-1α stabilization, a finding that was confirmed with independent SHARP1 siRNAs (Fig. 3d, e and Supplementary Fig. 21). In agreement with SHARP1 expression being regulated by TAp63α in TNBC cell lines (Supplementary Fig. 13), effective downregulation of endogenous HIF-1α protein could be observed in MDA-231 cells after transfecting an expression vector for TAp63α, or after knockdown of mutant p53 (Supplementary Fig. 22).

This raised the question of how SHARP1 impinges on HIF protein levels. Although oxygen-dependent or pVHL-mediated degradation of HIF-1α or HIF-2α has received considerable attention[7], it is also clear that HIFs are unstable proteins degraded by the proteasome even under hypoxic conditions[19,20]. However, this aspect of HIF regulation has remained only partially understood. We therefore tested to see whether the proteasome is required for the effects of SHARP1. Treatments of MDA-231 or HEK293T cells with proteasome inhibitors effectively opposed SHARP1-mediated HIF-1α degradation, resulting in accumulation of unmodified and polyubiquitinated HIF-1α isoforms (Fig. 3f and Supplementary Fig. 23). Notably, SHARP1 was able to downregulate the levels of wild-type HIF-1α

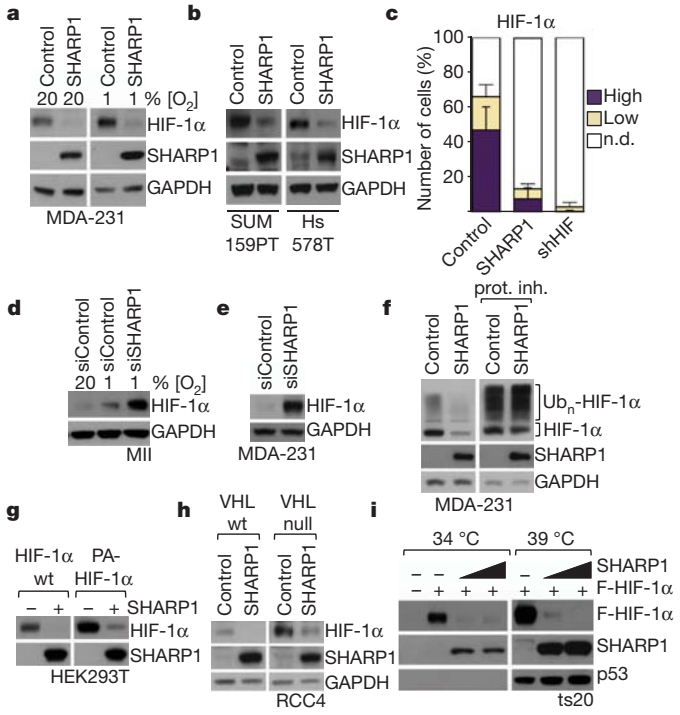


Figure 3 | SHARP1 promotes HIF-1α proteasomal degradation independently of oxygen levels and pVHL. **a**, Western blot analysis of control and SHARP1-overexpressing MDA-231 cells cultured at the indicated oxygen levels for 24 h. **b**, SHARP1 overexpression downregulates HIF-1α protein levels in two additional TNBC cell lines (SUM159 and Hs578T). **c**, Quantification of fluorescent immunohistochemistry of HIF-1α protein in tumours arising from orthotopically injected control and SHARP1-overexpressing MDA-231 cells (n = 7 tumours) (see Supplementary Fig. 20 for representative images). Error bars represent mean and s.e.m. **d**, **e**, Depletion of SHARP1 increases HIF-1α stabilization in MII cells (**d**, after 6 h of hypoxia) and MDA-231 cells (**e**, cultivated in normoxia) (for similar results obtained with an independent SHARP1-siRNA and for control of knockdown, see Supplementary Fig. 21). **f**, Downregulation of HIF-1α protein levels by SHARP1 in MDA-231 cells depends on the proteasome. Prot. Inh., proteasome inhibitors; Ub_n-HIF-1α, polyubiquitinated HIF-1α. **g**, Raising SHARP1 leads to reduction of wild-type (wt) HIF-1α or PA-HIF-1α levels. **h**, Western blot analyses of cell lysates from RCC4 cells, either null for pVHL or pVHL-reconstituted, stably expressing SHARP1 or empty vector as control. **i**, BALB/c 3T3-derived ts20 cells (BALB/c3T3ts20), bearing a temperature-sensitive (ts) ubiquitin conjugating enzyme (E1) mutant, were transfected with Flag–HIF-1α (F-HIF-1α), alone or with increasing doses of SHARP1. Cells were cultured at 34 °C and shifted at 39 °C to inactivate E1. Western blot analysis of p53 ensures the efficient inhibition of ubiquitin-dependent pathways[24]. In case of vertical slicing of blots of the same cell line, samples were obtained from the same experiment and pictures derive from blots processed in parallel.

and of a hydroxylation-mutant, and therefore prolyl hydroxylase-insensitive, PA-HIF-1α (Fig. 3g), and could effectively operate in VHL-mutant renal cell carcinoma (RCC4) cells (Fig. 3h). Thus, SHARP1 curtails HIF-1α activity independently of the pVHL pathway.

To test whether ubiquitination is involved in SHARP1-mediated HIF-1α degradation, we used cells carrying a temperature-sensitive mutant of the E1 ubiquitin-activating enzyme (UBE1). SHARP1 is equally active in permissive (34 °C) and non-permissive (39 °C) conditions (Fig. 3i). As a positive control for this assay, mdm2-mediated degradation of p53 was inhibited in non-permissive conditions (Fig. 3i). This suggests that SHARP1 operates independently of HIF-1α ubiquitination. Consistent with this, overexpression of SHARP1 does not increase polyubiquitination of HIF-1α, which is different from the effects of transfected VHL (Supplementary Fig. 24).

For proteasomal-dependent degradation, we proposed as an alternative mechanism to ubiquitination that SHARP1 may present HIF-1α to the proteasome. The proteasome has previously been shown to bind and degrade specific short-lived proteins in an ubiquitin-independent manner[21,22], matching the effects of SHARP1. Interestingly, unmodified HIF-1α can directly bind the 20S α4 subunit of the proteasome[23]. Through co-immunoprecipitations, both HIF-1α and SHARP1 associate with the 20S proteasomal subunit (Fig. 4a).

To capture in living cells if SHARP1 is instrumental for HIF-1α recognition by the proteasome, we stabilized protein complexes by treating HEK293T and MDA-231 cells with the bifunctional and cell permeable crosslinker DSP (dithiobis(succinimidylpropionate)) before cell lysis and co-immunoprecipitation. Expression of SHARP1 increased the association of HIF-1α with the 20S subunit (Fig. 4b and Supplementary Fig. 25). Given the requirement of SHARP1 for HIF-1α instability, we then tested the relevance of endogenous SHARP1 for the formation of the HIF-1α–20S complex. In co-immunoprecipitation assays, HIF-1α failed to efficiently associate to the 20S subunit in lysates of SHARP1-depleted cells, indicating that SHARP1 is essential for efficient recognition of HIF-1α by the 20S proteasome (Fig. 4c).

We next questioned whether the capacity to bind HIF and the proteasome is instrumental for SHARP1 function. To address this question we first dissected the structural requirements of SHARP1 to dock HIF-1α to the 20S by comparing different SHARP1 deletion constructs using co-immunoprecipitation assays. The amino-terminal basic helix–loop–helix (bHLH) domain of SHARP1 (Fig. 4d) was

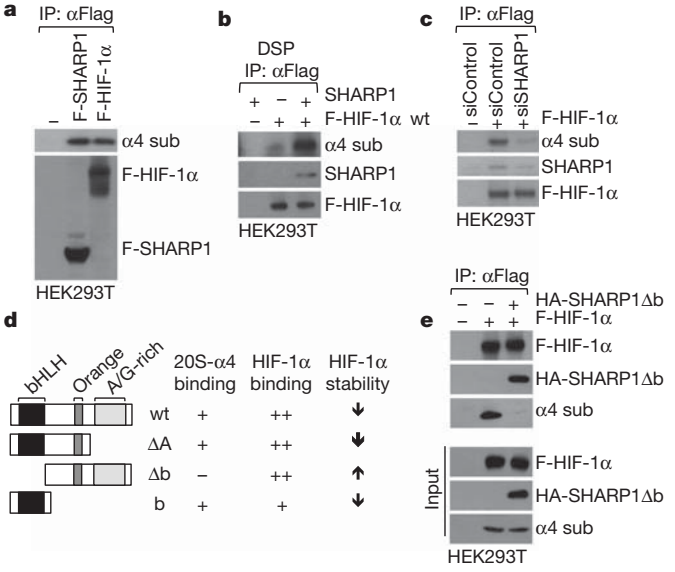


Figure 4 | SHARP1 promotes the interaction of HIF-1α with the proteasome. **a**, Co-immunoprecipitation of the α4 subunit of the 20S proteasome with overexpressed SHARP1 and HIF-1α. **b**, HEK293T cells expressing Flag-tagged HIF-1α alone or together with untagged SHARP1, were incubated with the cell-permeable DSP crosslinker before collection. Extracts were subjected to anti-Flag immunoprecipitation (IP), and co-precipitated endogenous α4 subunit of the 20S proteasome was detected by immunoblotting after de-crosslinking of the lysate. **c**, HEK293T cells were transfected with the indicated siRNAs and with Flag-tagged HIF-1α. After collection, extracts were subjected to anti-Flag immunoprecipitation; the co-precipitating endogenous 20S α4 proteasome subunit and endogenous SHARP1 proteins were detected by immunoblotting. **d**, Diagrams of the domains of SHARP1 and corresponding deletion mutants (ΔA, Δb and b) used in co-immunoprecipitation experiments with HIF-1α and the α4 subunit of the 20S proteasome, and a summary of the results (experimental data are shown in Supplementary Figs 26 and 27). Plus and minus symbols, presence or absence of binding of the different SHARP1 mutants with the 20S α4 subunits or HIF-1α. Arrows, effect of the SHARP1 mutant (increasing or decreasing HIF-1α stability). **e**, A SHARP1 deletion lacking its bHLH domain inhibits the binding of HIF-1α to the α4 proteasome subunit.

required and sufficient for proteasomal association, whereas for HIF association, the bHLH is sufficient but not essential (Fig. 4d and Supplementary Figs 26 and 27). The sole bHLH domain is capable of triggering HIF-1α and HIF-2α instability in a proteasome-dependent manner, thus recapitulating the effects of full-length SHARP1 (Fig. 4d and Supplementary Figs 28 and 29). Interestingly, a bHLH-deleted version of SHARP1 (SHARP1-Δb), retained HIF-1α association but lost proteasomal recognition (Fig. 4d and Supplementary Figs 26 and 27). Accordingly, expression of SHARP1-Δb led to HIF stabilization (Supplementary Figs 28 and 29) by preventing the binding of HIF-1α to the proteasome (Fig. 4e, as assayed by anti-HIF-1α co-immunoprecipitation). We conclude from these results that SHARP1 needs to associate with both HIFs and the proteasome to cause HIF degradation.

In the present work, we have presented a mechanistic link between SHARP1 and HIF activities, and we have provided clinical and functional evidence suggesting that this pathway is exploited in aggressive TNBC. A notable result was the identification of SHARP1 as an essential cellular determinant of the intrinsic instability of HIF proteins. SHARP1 acts in both normoxic and hypoxic cells, and irrespective of pVHL or ubiquitination pathways. We propose that, acting in parallel to oxygen-dependent mechanisms for HIFs degradation, elevated SHARP1 levels curtail the effects of hypoxic as well as oncogenic HIF stabilization in breast cancer. HIF activation represents a final common event in cancer pathogenesis in a variety of tumours; as such, the identification of SHARP1 uncovers the possibility of manipulating HIF-induced tumour progression.

METHODS SUMMARY

For trans-well migration assays, MDA-MB-231 cells were transfected with the indicated siRNAs, starved for 48 h in medium without serum and then plated on polyester (PET) inserts. The medium in the top and bottom chambers was supplemented with the TGFβ receptor inhibitor SB505124 (10 μM) or TGFβ1 (5 ng ml^{-1}) to induce cell migration. For protein–protein interactions in live cells between HIF-1α and the 20S α4 subunit, cells were treated with the cell-permeable crosslinker DSP (dithiobis-(succinimidyl propionate)) (2.5 mM) before harvesting. For ubiquitination assays with 3T3-ts20 cells, these cells were incubated at 39 °C for 8–12 h, transfected and then cultured at 39 °C for an additional 48 h. For all hypoxic treatments, cells were kept in a hypoxic cabinet and cultured at 1% oxygen level for the indicated times (for a complete description of methods and protocols, please see the Supplementary Information).

Received 6 April 2011; accepted 1 May 2012.
Published online 8 July 2012.

1. Elias, A. D. Triple-negative breast cancer: a short review. *Am. J. Clin. Oncol.* **33,** 637–645 (2010).
2. Di Cosimo, S. & Baselga, J. Management of breast cancer with targeted agents: importance of heterogeneity. *Nature Rev. Clin. Oncol.* **7,** 139–147 (2010).
3. Adorno, M. *et al.* A Mutant-p53/Smad complex opposes p63 to empower TGFβ-induced metastasis. *Cell* **137,** 87–98 (2009).
4. Su, X. *et al.* TAp63 suppresses metastasis through coordinate regulation of Dicer and miRNAs. *Nature* **467,** 986–990 (2010).
5. Muller, P. A. *et al.* Mutant p53 drives invasion by promoting integrin recycling. *Cell* **139,** 1327–1341 (2009).
6. Gordan, J. D. & Simon, M. C. Hypoxia-inducible factors: central regulators of the tumor phenotype. *Curr. Opin. Genet. Dev.* **17,** 71–77 (2007).
7. Kaelin, W. G. Jr & Ratcliffe, P. J. Oxygen sensing by metazoans: the central role of the HIF hydroxylase pathway. *Mol. Cell* **30,** 393–402 (2008).
8. Semenza, G. L. Defining the role of hypoxia-inducible factor 1 in cancer biology and therapeutics. *Oncogene* **29,** 625–634 (2010).
9. Helczynska, K. *et al.* Hypoxia-inducible factor-2α correlates to distant recurrence and poor outcome in invasive breast cancer. *Cancer Res.* **68,** 9212–9220 (2008).
10. Zhang, H. *et al.* HIF-1-dependent expression of angiopoietin-like 4 and L1CAM mediates vascular metastasis of hypoxic breast cancer cells to the lungs. *Oncogene* **31,** 1757–1770 (2011).
11. Lu, X. *et al. In vivo* dynamics and distinct functions of hypoxia in primary tumor growth and organotropic metastasis of breast cancer. *Cancer Res.* **70,** 3905–3914 (2010).
12. Sato, F. *et al.* Basic-helix-loop-helix (bHLH) transcription factor DEC2 negatively regulates vascular endothelial growth factor expression. *Genes Cells* **13,** 131–144 (2008).
13. Padua, D. *et al.* TGFβ primes breast tumors for lung metastasis seeding through angiopoietin-like 4. *Cell* **133,** 66–77 (2008).
14. Wong, C. C. *et al.* Hypoxia-inducible factor 1 is a master regulator of breast cancer metastatic niche formation. *Proc. Natl Acad. Sci. USA* **108,** 16369–16374 (2011).
15. Buffa, F. M. *et al.* microRNA-associated progression pathways and potential therapeutic targets identified by integrated mRNA and microRNA expression profiling in breast cancer. *Cancer Res.* **71,** 5635–5645 (2011).
16. Vleugel, M. M. *et al.* Differential prognostic impact of hypoxia induced and diffuse HIF-1α expression in invasive breast cancer. *J. Clin. Pathol.* **58,** 172–177 (2005).
17. Girardini, J. E. *et al.* A Pin1/mutant p53 axis promotes aggressiveness in breast cancer. *Cancer Cell* **20,** 79–91 (2011).
18. Strano, S. *et al.* Physical interaction with human tumor-derived p53 mutants inhibits p63 activities. *J. Biol. Chem.* **277,** 18817–18826 (2002).
19. Uchida, T. *et al.* Prolonged hypoxia differentially regulates hypoxia-inducible factor (HIF)-1α and HIF-2α expression in lung epithelial cells: implication of natural antisense HIF-1alpha. *J. Biol. Chem.* **279,** 14871–14878 (2004).
20. Kong, X., Alvarez-Castelao, B., Lin, Z., Castano, J. G. & Caro, J. Constitutive/hypoxic degradation of HIF-α proteins by the proteasome is independent of von Hippel Lindau protein ubiquitylation and the transactivation activity of the protein. *J. Biol. Chem.* **282,** 15498–15505 (2007).
21. Sdek, P. *et al.* MDM2 promotes proteasome-dependent ubiquitin-independent degradation of retinoblastoma protein. *Mol. Cell* **20,** 699–708 (2005).
22. Asher, G., Tsvetkov, P., Kahana, C. & Shaul, Y. A mechanism of ubiquitin-independent proteasomal degradation of the tumor suppressors p53 and p73. *Genes Dev.* **19,** 316–321 (2005).
23. Cho, S. *et al.* Binding and regulation of HIF-1α by a subunit of the proteasome complex, PSMA7. *FEBS Lett.* **498,** 62–66 (2001).
24. Chowdary, D. R., Dermody, J. J., Jha, K. K. & Ozer, H. L. Accumulation of p53 in a mutant cell line defective in the ubiquitin pathway. *Mol. Cell. Biol.* **14,** 1997–2003 (1994).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank O. Wessely, G. Del Sal, C. Oakman and A. Di Leo for comments; and W. Kaelin, C. Borner, S. Libutti, Y. Maeda and Y. Kato for gifts of reagents. M.M. is a recipient of a FIRC (Federazione Italiana Ricerca Cancro) fellowship. We are in debt to E. Tagliafico and the Modena Affimetrix platform team for help with microarrays. E.E. is a recipient of a Cariparo PhD fellowship. M.M. was a recipient of an AIRC (Italian Association for Cancer Research) fellowship. This work is supported by a Young Italian Researchers grant of the Italian Welfare Ministry and an AIRC (Associazione Italiana per la Ricerca sul Cancro) MFAG grant to M.C.; a Fondazione Città della Speranza Grant, and MIUR (Ministero dell'Istruzione dell'Università e della Ricerca Italia) and PRIN grants to G.B; and an AIRC Principal Investigator grant, an AIRC Special Program Molecular Clinical Oncology '5 per mille' grant, an HSFP grant, a University of Padua Strategic grant, an IIT Excellence grant, a CNR–Miur Epigenetics Flagship project, and a Comitato Promotore Telethon grant to S.P.

Author Contributions M.M., M.C. and S.P. designed research, and M.M., E.E, M.F., M.C., E.R., G.B., G.L. and F.Z. performed experiments. M.C., S.B. and M.F. performed bioinformatics analysis, and A.R. and A.P. helped with assays in mice and tumour pathology. M.M., M.C. and S.P. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.P. (piccolo@bio.unipd.it).

LETTER

doi:10.1038/nature11169

Structure of the immature retroviral capsid at 8 Å resolution by cryo-electron microscopy

Tanmay A. M. Bharat[1], Norman E. Davey[1]*, Pavel Ulbrich[2]*, James D. Riches[1], Alex de Marco[1], Michaela Rumlova[3], Carsten Sachse[1], Tomas Ruml[2] & John A. G. Briggs[1]

The assembly of retroviruses such as HIV-1 is driven by oligomerization of their major structural protein, Gag. Gag is a multidomain polyprotein including three conserved folded domains: MA (matrix), CA (capsid) and NC (nucleocapsid)[1]. Assembly of an infectious virion proceeds in two stages[2]. In the first stage, Gag oligomerization into a hexameric protein lattice leads to the formation of an incomplete, roughly spherical protein shell that buds through the plasma membrane of the infected cell to release an enveloped immature virus particle. In the second stage, cleavage of Gag by the viral protease leads to rearrangement of the particle interior, converting the non-infectious immature virus particle into a mature infectious virion. The immature Gag shell acts as the pivotal intermediate in assembly and is a potential target for anti-retroviral drugs both in inhibiting virus assembly and in disrupting virus maturation[3]. However, detailed structural information on the immature Gag shell has not previously been available. For this reason it is unclear what protein conformations and interfaces mediate the interactions between domains and therefore the assembly of retrovirus particles, and what structural transitions are associated with retrovirus maturation. Here we solve the structure of the immature retroviral Gag shell from Mason–Pfizer monkey virus by combining cryo-electron microscopy and tomography. The 8-Å resolution structure permits the derivation of a pseudo-atomic model of CA in the immature retrovirus, which defines the protein interfaces mediating retrovirus assembly. We show that transition of an immature retrovirus into its mature infectious form involves marked rotations and translations of CA domains, that the roles of the amino-terminal and carboxy-terminal domains of CA in assembling the immature and mature hexameric lattices are exchanged, and that the CA interactions that stabilize the immature and mature viruses are almost completely distinct.

Within the immature virus particle the Gag polyprotein is arranged radially, with the N-terminal MA domain at the viral membrane, and the NC domain pointing towards the centre of the particle (Supplementary Fig. 1). Between MA and NC, interactions involving CA and residues immediately downstream of CA are the primary mediators of Gag assembly[1,2]. Gag constructs can be assembled *in vitro* to form roughly spherical immature virus-like particles (VLPs)[4,5]. NC–RNA–NC interactions support oligomerization but can be replaced *in vitro* with a non-specific protein interaction domain such as a leucine zipper[6,7]. Proteolytic cleavage leads to disassembly of the immature Gag shell and formation of the mature virion, which has a very different appearance: MA remains underneath the viral membrane, and the NC–RNA complex is condensed in the centre of the particle surrounded by a CA core, which can be tubular, conical or polyhedral depending on the virus. In both the immature Gag shell and the mature core, CA assembles to form a curved hexameric protein lattice, but the spacing and arrangement of the proteins in the two lattices are different[8–11].

X-ray crystallography and NMR-derived structures of individual Gag domains have shown that both of the folded CA domains are predominantly α-helical: the N-terminal CA domain (CA-NTD) contains seven α-helices, and the C-terminal domain (CA-CTD) contains four. Their structures are highly conserved across all retroviruses (Supplementary Fig. 2). Crystallographic and electron microscopy studies of reconstituted assemblies of viral proteins have provided a detailed structural view of the arrangement of the CA domains within the mature capsid core in HIV-1 and RSV[12–15]. No such information is available for the immature retroviral Gag shell: studies of it have proved more challenging for two reasons. First, Gag is a flexible polyprotein that cannot be crystallized whole and adopts the immature-like conformation only on assembly into the lattice. Second, the resulting lattice is curved and flexible; it has therefore not been accessible to high-resolution structural techniques. Low-resolution studies of immature HIV-1 particles and VLPs by cryo-electron tomography (cryo-ET)[16,17] have revealed that the CA lattice has an outer CA-NTD layer arranged as hexameric rings around large holes, below which CA-CTDs form two-fold symmetrical densities, linking between hexamers. The same characteristic arrangement of CA domains is also seen in roughly spherical immature VLPs of other retroviruses[18]. Outside the CA region, Gag adopts different structures in different retroviruses[18].

A Mason–Pfizer monkey virus (M-PMV) truncated Gag construct consisting of CA and NC with the N-terminal proline of CA deleted (ΔPro M-PMV CANC), can be assembled together with oligonucleotides to form roughly spherical immature VLPs[18,19]. We observed that under certain assembly conditions the spheres were mixed with broad tubes of similar diameter to the spheres (Fig. 1a and Supplementary Table 1). In some cases the tubes were continuous with partial spherical particles (Supplementary Fig. 3a). We applied cryo-ET and image-processing techniques to understand the structure of the Gag shell in the tubes. Comparison of radial density profiles, power spectra of the lattices, and subtomogram averaging reconstructions confirmed that all analysed tubes share the same characteristic immature virus-like arrangement of Gag as the spheres (Supplementary Fig. 3).

We wished to generate a high-resolution three-dimensional reconstruction of the tubes by using real-space helical reconstruction[20]. The application of this method was limited, because the range of different tube diameters (Supplementary Table 1) prevented the unambiguous determination of helical symmetry parameters from two-dimensional images[21]. We therefore combined cryo-electron microscopy (cryo-EM) and cryo-ET methods to derive the average structure of the tubes (Supplementary Fig. 4a). For each tube a two-dimensional image was collected on film, and then cryo-ET data were collected for the same tube. By using subtomogram averaging, the arrangement of the hexamers on each individual tube was determined[16]. On the basis of this information, distorted tubes were identified and excluded, and the

[1]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. [2]Department of Biochemistry and Microbiology, Institute of Chemical Technology, Prague, Technicka 3, 166 28 Prague, Czech Republic. [3]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of Czech Republic, v.v.i., Flemingovo nám. 2, 166 10 Prague, Czech Republic.
*These authors contributed equally to this work.

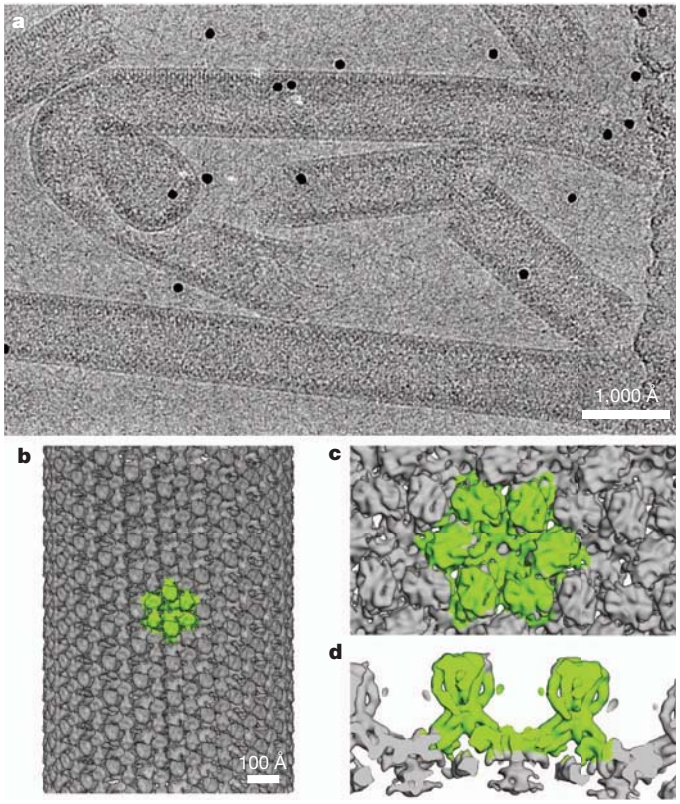


Figure 1 | Cryo-EM reconstruction of M-PMV CANC tubes. **a**, A representative cryo-EM image of the M-PMV CANC tubes. Protein density is black. **b**, Real-space helical reconstruction of one single tube. One hexamer of Gag dimers is highlighted in green. **c**, Reconstruction of the M-PMV CANC tubes at 9 Å resolution by averaging of the pseudo-hexameric asymmetric unit from 25 tubes with 9 different symmetries (Supplementary Table 1). The isosurface has been thresholded at 2.0σ from the mean. **d**, The same reconstruction in an orthogonal orientation.

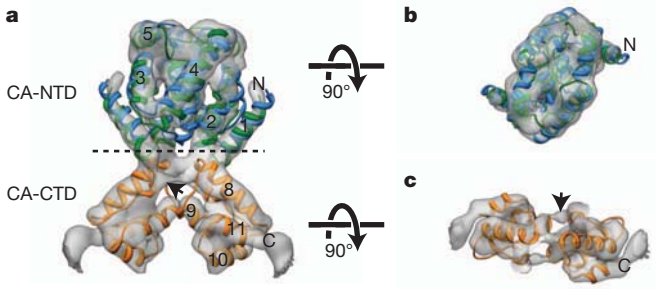


Figure 2 | Fitting of known atomic structures into the cryo-EM map. **a**, A fit of the M-PMV CA-NTD NMR structure (green ribbon) into the 8-Å cryo-EM map. The CA density from one Gag dimer is shown as an isosurface thresholded at 3.5σ from the mean. The conserved HIV-1 CA-NTD (blue) and CA-CTD (orange) domains have been additionally fitted into the density. Helix numbers and the N and C termini of CA are marked. **b**, **c**, Orthogonal orientations of the CA-NTD (**b**) and CA-CTD (**c**) domains from the same fit separated along the clipping plane shown in **a** as dotted lines. Cryo-EM density for the 7–8 linker is partly visible and is marked in **a** and **c** with a black arrow. See also Supplementary Movie 1.

helical symmetry parameters for each remaining tube were extracted. These symmetry parameters were refined and then used for helical reconstruction of the tube from the two-dimensional film image (Fig. 1b). In this way, three-dimensional reconstructions of 25 individual tubes with 9 different symmetries were generated (Supplementary Table 1 and Supplementary Methods). To combine the information from the individual tubes, the pseudo-hexameric asymmetric units from each of the individual tube reconstructions were extracted and were averaged in three dimensions to generate a reconstruction at 9 Å resolution (Fig. 1c, d and Supplementary Movie 1). Further averaging of the three independent copies of Gag within the asymmetric unit increased the resolution to 8 Å (Fig. 2 and Supplementary Fig. 5), from a total data set of 344,760 copies of Gag.

Densities for all 11 α-helices in the two CA domains are resolved in the cryo-EM reconstruction, allowing the unambiguous positioning of high-resolution atomic structures to generate a pseudo-atomic model of the immature CA lattice. Both CA-NTD and CA-CTD domains adopt the same fold in the assembled lattice as in the high-resolution structures of isolated domains. The NMR structure of the M-PMV CA-NTD domain[22] is shown fitted into the cryo-EM density in Fig. 2. Because of the high degree of structural conservation, CA-NTD and CA-CTD domains from different retroviruses including HIV-1 also fit well into the cryo-EM density (Fig. 2 and Supplementary Fig. 5). Consistent with the general architecture suggested by previous low-resolution studies[16,17], the CA-NTD domain forms the outermost CA layer containing large holes on the pseudo-six-fold symmetry axes in the hexagonal lattice (Fig. 1c). The CA-CTD domains are located below the CA-NTD domains, where they link between the six-fold positions by forming a dimeric interaction at the two-fold symmetry axes in the lattice (Figs 2 and 3a, b (left)). Density for part of the long extended chain connecting helix 7 in CA-NTD to helix 8 in CA-CTD (the '7–8 linker') is visible in the map, near the base of helix 1, positioning the CA-NTD domains almost directly above the CA-CTD domains (Fig. 2 and Supplementary Movie 1). Immediately downstream of the fitted HIV-1 CA-CTD domain there is an additional kink in the cryo-EM density at a position of a proline residue in the M-PMV sequence (position 517), followed by a short structured density (the 'CA-CTD tail'). The CA-CTD tail is located in the vicinity of the loop between the 7–8 linker and helix 8 of CA-CTD (Supplementary Fig. 6a). This loop, together with the adjacent regions of the 7–8 linker and helix 8, comprises the most conserved region of CA, the major homology region[1]. Disassembly of the immature lattice requires proteolytic cleavage both upstream of helix 1 in CA-NTD and downstream of the CA-CTD tail[23]. Within the immature lattice, both helix 1 and the CA-CTD tail are in a position to interact directly with the 7–8 linker and potentially to communicate structural changes between domains.

The fitted cryo-EM reconstruction reveals the CA interactions within the immature lattice. The CA-NTD domain does not form direct intra-hexamer interactions; instead the CA-CTD forms intra-hexameric interactions, which would bring HIV residues T348 and Q351 of helix 11 into proximity with residues G288, P289 and K290 in the major homology region of the adjacent CA-CTD in the ring (Fig. 3c (green) and Supplementary Movie 1). There are no substantial CA-NTD–CA-CTD interactions in the immature lattice, which is consistent with predictions from ^{1}H–^{2}H exchange experiments[24]. Below the CA-CTD tail, in the CA–NC spacer region, a ring of protein links the six Gag monomers in the hexamer (Supplementary Fig. 6b). There are no atomic structures available for this part of Gag, and its structure is not conserved between retroviruses[18].

The hexamers are linked to one another by large inter-hexamer dimerization interfaces between CA-NTD domains, and between CA-CTD domains. The CA-NTD domain forms a homodimeric interaction through residues in helices 4, 5, 6 and 7 of CA (Fig. 3d, magenta). This CA-NTD interaction differs significantly from those proposed on the basis of low-resolution cryo-ET data[16,17]. The CA-NTD dimer is linked to adjacent dimers by a smaller trimeric interaction interface involving the N terminus of helix 4. Consistent with predictions from low-resolution tomographic data[16], the immature CA-CTD forms a homodimer that is narrow and elongated. It resembles the dimeric form found in crystals of HIV-1 CA-CTD in the presence of CAI, a peptide assembly inhibitor[25]. This dimeric form is also observed in a Y301A mutant of the HIV-1 CA-CTD, which is competent for immature virus assembly but not for mature core

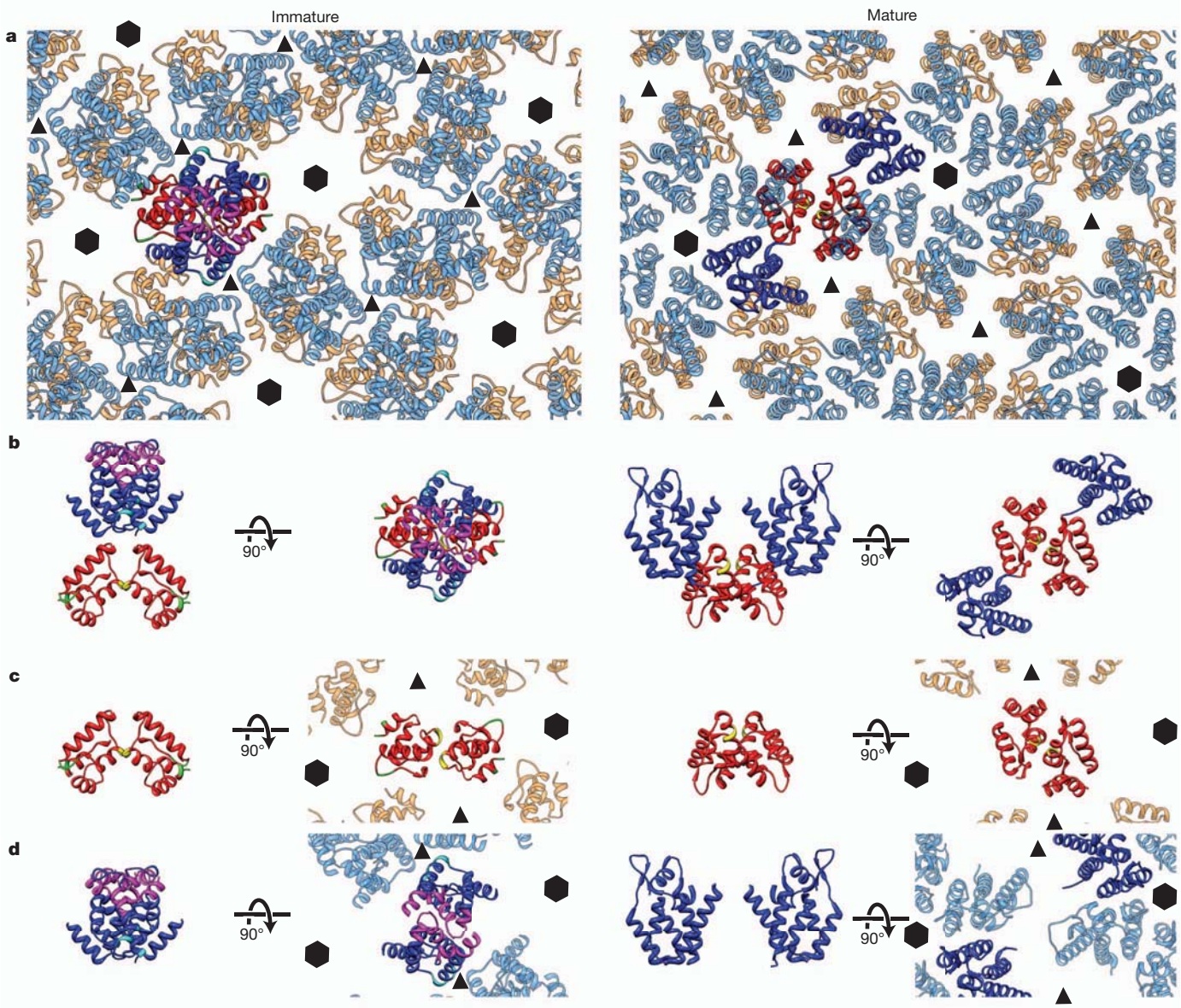


Figure 3 | Comparison of the arrangement of CA domains in the mature and immature retrovirus lattices. **a**, The arrangement of CA domains in the immature (left) and the mature (right) lattices is shown. Fitted HIV-1 CA-NTD domains are blue and light blue, and HIV-1 CA-CTD domains are red and orange. A schematic view of the structures is shown in Supplementary Fig. 1, and the transition between the two lattices is illustrated in Supplementary Movie 2. The six-fold and three-fold rotational symmetry axes are highlighted with black hexagons and triangles, respectively. **b**, One immature (left) and mature (right) Gag dimer is shown in two orthogonal orientations. **c**, The CA-CTD dimers in both lattices are shown in the corresponding orientations. Residues that interact in both immature and mature lattices—W316 and M317 in helix 9 of CA-CTD—are highlighted in yellow. CA-CTD residues mediating interfaces around the immature hexamer are in light green (left). **d**, One immature CA-NTD dimer is shown with dimeric and trimeric interface residues highlighted in magenta and cyan, respectively (left). The dimeric interface is absent from the mature lattice (right).

assembly (Supplementary Fig. 7)[26]. The homodimeric CA-CTD interaction is mediated by helix 9. Mutations of Q311, V313, W316 or M317 in helix 9 of HIV-1 CA-CTD are known to prevent assembly of the Gag lattice[27,28]. The length of the cryo-EM density corresponding to helix 9 indicates that its N-terminal residues (HIV-1 Gag 311–315) do not form a structured α-helix in the immature lattice structure (Supplementary Fig. 7a, b). The same residues do not form a helix in several crystal forms of HIV-1 CA-CTD (for example 3MGE, 3H47 and some chains of 3H4E).

Assembly of the immature particle, and proteolytic cleavage of the immature Gag polyprotein to trigger maturation, are both required for infectivity. The immature Gag shell is therefore an important potential target for antiretroviral drugs[3]. The immature retrovirus particle is an 'assembly machine' that selects and incorporates the necessary viral and cellular components from within the crowded cytoplasm of the infected cell and initiates membrane envelopment during budding. In contrast, assembly of the mature capsid core occurs within the shelter of the virus particle at high CA concentrations and results in a metastable core destined to undergo transport and regulated disassembly within the target cell. These contrasting functional requirements are both mediated by the assembly of CA into a hexameric lattice. We compared the arrangement of CA proteins in the immature and mature hexameric lattices and found that the transition between these functional requirements is achieved through a spectacular and conclusive structural change (Fig. 3 and Supplementary Movie 2). First, the relative position and orientation of the CA-CTD and CA-NTD domains change completely (Fig. 3b and Supplementary Fig. 1b, c). Second, the roles of the two domains in assembling the hexameric lattice are largely exchanged (Table 1). The immature hexamer is formed by CA-CTD–CA-CTD interactions and by residues downstream of CA,

Table 1 | Number of interface residues in the immature and mature retroviral lattices

Interaction	Immature		Mature	
	Intra-hexamer	Inter-hexamer	Intra-hexamer	Inter-hexamer
CA-NTD–CA-NTD	0	39	23	0
CA-NTD–CA-CTD	0	0	20	0
CA-CTD–CA-CTD	5	6	0	7

Numbers of interface residues (including loops) in both the mature and immature lattices are shown (Supplementary Information). Intra-hexameric CA-CTD–CA-CTD and inter-hexameric CA-NTD–CA-NTD contacts in the immature lattice are replaced by intra-hexameric CA-NTD–CA-CTD and CA-NTD–CA-NTD interactions in the mature lattice.

whereas the mature hexamer is formed by CA-NTD—CA-NTD interactions involving helices 1, 2 and 3 (Fig. 3a, d). In the immature lattice, CA-NTD–CA-NTD interfaces involving helices 4, 5, 6 and 7 link hexamers to one another, whereas in the mature lattice there are no inter-hexamer interactions involving CA-NTD (Fig. 3d). CA-NTD–CA-CTD interactions are present in the mature lattice but absent from the immature lattice (Fig. 3b). Third, no interfaces are conserved in both immature and mature lattices. Indeed, only two residues of HIV-1, the assembly-critical W316 and M317 positions in the CA-CTD helix 9 dimer interface, are within 8 Å of the same interaction partner in both immature and mature lattices (Fig. 3c, yellow), and the CA-CTD domains rotate by 119° about this interface on maturation. Fourth, the sets of residues that form interfaces before and after maturation are largely independent (Fig. 4 and Supplementary Table 2). Rigid-body domain movements resulting in new protein–protein interfaces also take place during the assembly of some icosahedral viruses including Dengue virus[29] and the bacteriophage HK97 (ref. 30). In these viruses, the protein shell retains its integrity during the movements. Protein movements during retroviral capsid maturation occur within the envelope and are accompanied by disassembly and reassembly of the protein lattice. Large-scale structural maturation processes allow the same proteins to assemble immature and mature virus particles with contrasting functions of assembly and entry. Further, they make different sets of residues available for interacting with cellular and viral proteins during immature virus assembly to those available during entry of the mature capsid into the target cell.

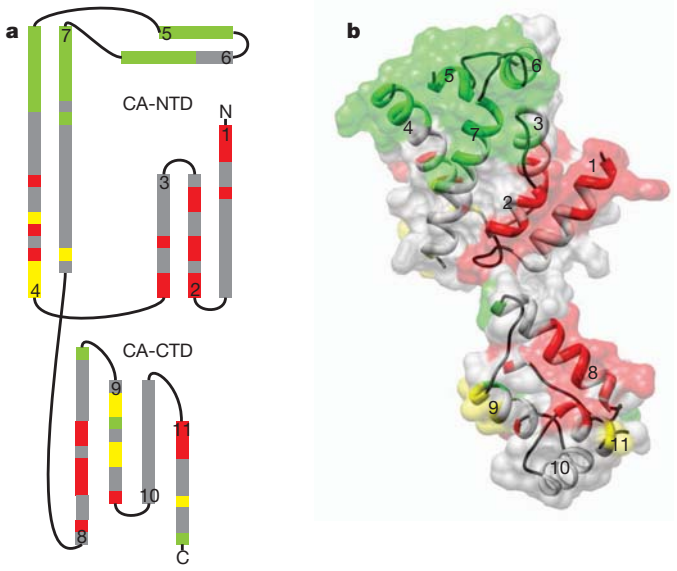


Figure 4 | Modularity of protein interfaces. **a**, A topological diagram of the 11 CA α-helices (numbered), with the HIV-1 sequence as a model. Residues that are in a position to form a CA–CA interface in the immature lattice, the mature lattice, or both are coloured green, red and yellow, respectively (Supplementary Information and Supplementary Table 2). Other helical residues are grey, and loops are shown as black lines. **b**, An immature HIV-1 Gag monomer is shown with residues highlighted using the same colour scheme as in **a** with a surface representation of the same protein overlaid.

METHODS SUMMARY

To assemble tubes, ΔPro M-PMV CANC protein[19] was mixed with λ-DNA or MS2-RNA in a buffer containing 60 mM dithiothreitol (DTT), 500 mM NaCl, 1 μM $ZnCl_2$ and 50 mM phosphate pH 7.5, and dialysed into buffer with 20 mM DTT, 100 mM NaCl, 50 mM Tris pH 7.8 at 25 °C. The resulting sample was mixed with a buffered solution of Protein A conjugated with 10-nm gold beads and vitrified by plunge-freezing in liquid ethane. Cryo-EM and cryo-ET were performed on a FEI Titan Krios operated at 300 kV at liquid nitrogen temperature. Images were collected onto Kodak SO-163 film, and digitized with a Zeiss SCAI scanner at 7 μm pixel step giving a final unbinned pixel size of 1.5 Å. Cryo-ET data for the same region were collected on a Gatan 2k charge-coupled device camera using a GIF2002 post-column energy filter at a pixel size of 4.9 Å. The helical parameters of each tube were extracted using cryo-ET and subtomogram averaging[16]. These parameters were used to apply real-space helical reconstruction techniques to obtain structures of individual tubes with different symmetries[20]. The pseudo-hexameric asymmetric units from all reconstructed tubes were extracted and averaged in three dimensions to obtain a final reconstruction of the immature CA lattice at about 9 Å and of the Gag dimer at about 8 Å. Atomic structures of retroviral CA domains were fitted into the cryo-EM densities as rigid bodies.

Received 28 February; accepted 27 April 2012.
Published online 3 June 2012.

1. Göttlinger, H. G. The HIV-1 assembly machine. *AIDS* **15** (Suppl. 5), S13–S20 (2001).
2. Briggs, J. A. & Kräusslich, H. G. The molecular architecture of HIV. *J. Mol. Biol.* **410,** 491–500 (2011).
3. Waheed, A. A. & Freed, E. O. HIV type 1 Gag as a target for antiviral therapy. *AIDS Res. Hum. Retroviruses* **28,** 54–75 (2012).
4. Gross, I., Hohenberg, H., Huckhagel, C. & Kräusslich, H. G. N-terminal extension of human immunodeficiency virus capsid protein converts the *in vitro* assembly phenotype from tubular to spherical particles. *J. Virol.* **72,** 4798–4810 (1998).
5. von Schwedler, U. K. *et al.* Proteolytic refolding of the HIV-1 capsid protein amino-terminus facilitates viral core assembly. *EMBO J.* **17,** 1555–1568 (1998).
6. Johnson, M. C., Scobie, H. M., Ma, Y. M. & Vogt, V. M. Nucleic acid-independent retrovirus assembly can be driven by dimerization. *J. Virol.* **76,** 11177–11185 (2002).
7. Accola, M. A., Strack, B. & Göttlinger, H. G. Efficient particle production by minimal Gag constructs which retain the carboxy-terminal domain of human immunodeficiency virus type 1 capsid-p2 and a late assembly domain. *J. Virol.* **74,** 5395–5402 (2000).
8. Yeager, M., Wilson-Kubalek, E. M., Weiner, S. G., Brown, P. O. & Rein, A. Supramolecular organization of immature and mature murine leukemia virus revealed by electron cryo-microscopy: implications for retroviral assembly mechanisms. *Proc. Natl Acad. Sci. USA* **95,** 7299–7304 (1998).
9. Li, S., Hill, C. P., Sundquist, W. I. & Finch, J. T. Image reconstructions of helical assemblies of the HIV-1 CA protein. *Nature* **407,** 409–413 (2000).
10. Briggs, J. A., Wilk, T., Welker, R., Kräusslich, H. G. & Fuller, S. D. Structural organization of authentic, mature HIV-1 virions and cores. *EMBO J.* **22,** 1707–1715 (2003).
11. Briggs, J. A. *et al.* The stoichiometry of Gag protein in HIV-1. *Nature Struct. Mol. Biol.* **11,** 672–675 (2004).
12. Ganser-Pornillos, B. K., Cheng, A. & Yeager, M. Structure of full-length HIV-1 CA: a model for the mature capsid lattice. *Cell* **131,** 70–79 (2007).
13. Pornillos, O. *et al.* X-ray structures of the hexameric building block of the HIV capsid. *Cell* **137,** 1282–1292 (2009).
14. Pornillos, O., Ganser-Pornillos, B. K. & Yeager, M. Atomic-level modelling of the HIV capsid. *Nature* **469,** 424–427 (2011).
15. Cardone, G., Purdy, J. G., Cheng, N., Craven, R. C. & Steven, A. C. Visualization of a missing link in retrovirus capsid assembly. *Nature* **457,** 694–698 (2009).
16. Briggs, J. A. *et al.* Structure and assembly of immature HIV. *Proc. Natl Acad. Sci. USA* **106,** 11090–11095 (2009).
17. Wright, E. R. *et al.* Electron cryotomography of immature HIV-1 virions reveals the structure of the CA and SP1 Gag shells. *EMBO J.* **26,** 2218–2226 (2007).
18. de Marco, A. *et al.* Conserved and variable features of Gag structure and arrangement in immature retrovirus particles. *J. Virol.* **84,** 11729–11736 (2010).
19. Ulbrich, P. *et al.* Distinct roles for nucleic acid in *in vitro* assembly of purified Mason-Pfizer monkey virus CANC proteins. *J. Virol.* **80,** 7089–7099 (2006).
20. Sachse, C. *et al.* High-resolution electron microscopy of helical specimens: a fresh look at tobacco mosaic virus. *J. Mol. Biol.* **371,** 812–835 (2007).
21. Egelman, E. H. Reconstruction of helical filaments and tubes. *Methods Enzymol.* **482,** 167–183 (2010).
22. Macek, P. *et al.* NMR structure of the N-terminal domain of capsid protein from the Mason-Pfizer monkey virus. *J. Mol. Biol.* **392,** 100–114 (2009).
23. de Marco, A. *et al.* Structural analysis of HIV-1 maturation using cryo-electron tomography. *PLoS Pathog.* **6,** e1001215 (2010).
24. Lanman, J. *et al.* Key interactions in HIV-1 maturation identified by hydrogen–deuterium exchange. *Nature Struct. Mol. Biol.* **11,** 676–677 (2004).
25. Ternois, F., Sticht, J., Duquerroy, S., Kräusslich, H. G. & Rey, F. A. The HIV-1 capsid protein C-terminal domain in complex with a virus assembly inhibitor. *Nature Struct. Mol. Biol.* **12,** 678–682 (2005).

26. Bartonova, V. *et al.* Residues in the HIV-1 capsid assembly inhibitor binding site are essential for maintaining the assembly-competent quaternary structure of the capsid protein. *J. Biol. Chem.* **283,** 32024–32033 (2008).
27. Chu, H. H., Chang, Y. F. & Wang, C. T. Mutations in the alpha-helix directly C-terminal to the major homology region of human immunodeficiency virus type 1 capsid protein disrupt Gag multimerization and markedly impair virus particle production. *J. Biomed. Sci.* **13,** 645–656 (2006).
28. von Schwedler, U. K., Stray, K. M., Garrus, J. E. & Sundquist, W. I. Functional surfaces of the human immunodeficiency virus type 1 capsid protein. *J. Virol.* **77,** 5439–5450 (2003).
29. Yu, I. M. *et al.* Structure of the immature dengue virus at low pH primes proteolytic maturation. *Science* **319,** 1834–1837 (2008).
30. Conway, J. F. *et al.* Virus maturation involving large subunit rotations and local refolding. *Science* **292,** 744–748 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This study was technically supported by the use of the European Molecular Biology Laboratory Information Technology Service unit. This work was partly funded by a grant from the Deutsche Forschungsgemeinschaft within SPP1175 to J.A.G.B. and by grants P302/12/1895 and 204/09/1388 from the Czech Science foundation to T.R. and M.R.

Author Contributions T.A.M.B., P.U., M.R., T.R. and J.A.G.B. designed the research. T.A.M.B. and P.U. prepared samples for electron microscopy. T.A.M.B. and J.D.R. collected cryo-EM data. T.A.M.B., J.D.R., A.D.M. and J.A.G.B. analysed cryo-ET data. C.S. supported helical image-processing techniques. T.A.M.B. and J.A.G.B. developed and applied the variable-symmetry helical reconstruction methodology. T.A.M.B., N.D., P.U., M.R., C.S., T.R. and J.A.G.B. analysed fitted pseudo-atomic models. T.A.M.B. and J.A.G.B. wrote the paper with support from all the authors.

Author Information Cryo-EM structural data have been deposited at the EMDB under accession numbers EMD-2089 and EMD-2090, and at the Protein Data Bank under accession numbers 4ard and 4arg. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.A.G.B. (briggs@embl.de).

LETTER

doi:10.1038/nature11172

Heterogeneous pathways and timing of factor departure during translation initiation

Albert Tsai[1,2], Alexey Petrov[1], R. Andrew Marshall[1,3], Jonas Korlach[4], Sotaro Uemura[1,5] & Joseph D. Puglisi[1,6]

The initiation of translation establishes the reading frame for protein synthesis and is a key point of regulation[1]. Initiation involves factor-driven assembly at a start codon of a messenger RNA of an elongation-competent 70S ribosomal particle (in bacteria) from separated 30S and 50S subunits and initiator transfer RNA. Here we establish in *Escherichia coli*, using direct single-molecule tracking, the timing of initiator tRNA, initiation factor 2 (IF2; encoded by *infB*) and 50S subunit joining during initiation. Our results show multiple pathways to initiation, with orders of arrival of tRNA and IF2 dependent on factor concentration and composition. IF2 accelerates 50S subunit joining and stabilizes the assembled 70S complex. Transition to elongation is gated by the departure of IF2 after GTP hydrolysis, allowing efficient arrival of elongator tRNAs to the second codon presented in the aminoacyl-tRNA binding site (A site). These experiments highlight the power of single-molecule approaches to delineate mechanisms in complex multicomponent systems.

Initiation is a key point of regulation of gene expression before the ribosome is committed to the energy-intensive process of synthesizing a full protein[1]. Protein factors guide and regulate initiation; three initiation factors, IF1 (encoded by *infA*), IF2 and IF3 (encoded by *infC*), are required for viability in bacteria, whereas a far larger complement of factors exists in eukaryotes.

Although the mechanism and overall kinetics of translation initiation in bacteria have been delineated over the past two decades[2], fundamental questions remain. The possible configurations that this multifactor system can adopt challenge traditional biophysical methods. The timings of individual factor and tRNA assembly on the ribosome, their coordination with each other, and the subsequent factor dissociation that allows elongation are not known. Translation initiation may follow a linear mechanism, or branch through multiple parallel pathways. We apply real-time single-molecule methods to track directly the dynamics of translation initiation in the model *E. coli* system. We determined the relative timing of initiator tRNA, IF2 and subunit binding, and showed how IF2 and GTP hydrolysis control the transition into elongation. Our data demonstrate that intermediate and late steps in initiation occur through heterogeneous pathways. The overall initiation rates and efficiency depend on the initiation pathway, whose selection is guided by initiation factors.

Single-molecule fluorescence experiments allow direct observation of dynamics in complex biological systems. To monitor single-molecule fluorescence at high (0.1–5 μM) concentrations of free dye-labelled biomolecules, optical confinement was achieved using zero-mode waveguides (ZMWs)[3] (Supplementary Fig. 1). We recently demonstrated the power of this approach by tracking the real-time dynamics of tRNA transit through the ribosome during elongation[4]. After conducting control experiments to verify the functionality of our dye-labelled biomolecules (Supplementary Text 1 and Supplementary Fig. 2), we broaden this method to follow tRNAs, protein factors and ribosomal subunits directly during initiation and transition into elongation (Fig. 1a).

Although recent experiments suggested that IF2 and initiator tRNA ($fMet\text{-}tRNA^{fMet}$) bind sequentially to the small subunit in the formation of a 30S pre-initiation complex (30S PIC)[5], IF2 with GTP bound

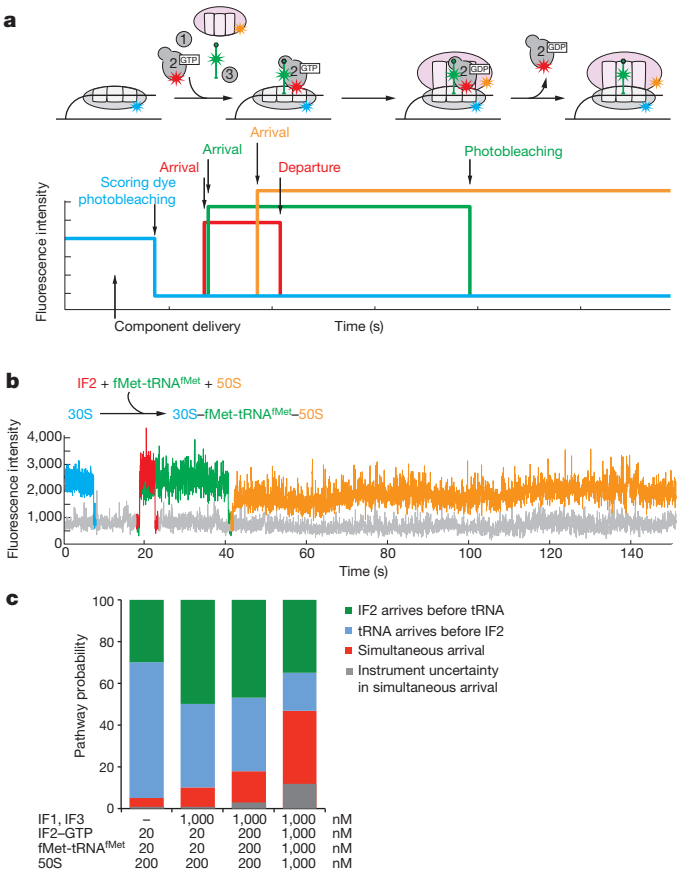


Figure 1 | Pathways leading to 30S PIC formation. a, Single dye-labelled 30S complexes were immobilized on the bottom of ZMW wells and scored by fluorescence (see Methods). Dye-labelled initiation factors, tRNAs and 50S subunits were delivered at $t = 7$ s in all experiments. The appearance of fluorescence signals indicates arrival of the labelled molecules. Fluorescence signal disappears either due to dye-labelled molecule departure or photobleaching. **b**, Productive initiation events were identified by stable 50S arrival (see Methods). The order of arrival was determined by the sequence of the fluorescent pulses. Grey portions of the traces represent fluorescence background. **c**, The ratios of possible 30S PIC formation pathways at different ligand concentrations were measured and plotted. See Methods for an explanation of instrument uncertainty. From left to right, $n = 86$, $n = 51$, $n = 79$ and $n = 52$.

[1]Department of Structural Biology, Stanford University School of Medicine, Stanford, California 94305-5126, USA. [2]Department of Applied Physics, Stanford University, Stanford, California 94305-4090, USA. [3]McKinsey & Company - Silicon Valley, 3705A Hansen Way, Palo Alto, California 94304, USA. [4]Pacific Biosciences, 1380 Willow Rd, Menlo Park, California 94025, USA. [5]Omics Science Center, RIKEN Yokohama Institute, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa 230-0045, Japan. [6]Stanford Magnetic Resonance Laboratory, Stanford University School of Medicine, Stanford, California 94305-5126, USA.

(IF2–GTP) also forms a weak complex with the tRNA (dissociation constant $K_d \approx 1\ \mu M$)[6,7], potentially allowing both to bind simultaneously. We determined whether IF2 and tRNA binding is simultaneous, sequential or random by delivering a mixture of Cy3-labelled initiator tRNA (fMet-(Cy3)tRNA^{fMet}), Cy5-labelled IF2 (Cy5–IF2) and Cy3.5-labelled large subunits (Cy3.5–50S) to immobilized 30S subunits labelled with Alexa488 at 20–1,000 nM of each reagent (see Methods). The appearance of a stable 50S signal ($t > 10$ s) was used to identify productive tRNA- and IF2-binding events. The relative timing of IF2 and initiator tRNA arrival to the ribosome was determined by single-molecule analysis (Fig. 1b).

At low concentrations (20 nM each) of IF2 and the initiator tRNA, tRNA arrives first in 65% of the initiation events, IF2 arrives first in 30%, and only 5% show simultaneous arrival of both molecules (Fig. 1c and Supplementary Fig. 3). Addition of IF1 and IF3 shifts the arrival order, with 50% of ribosomes having IF2 arrive before initiator tRNA, 40% having initiator tRNA before IF2, and 10% showing simultaneous arrival. This is consistent with IF1 and IF3 destabilizing initiator tRNA in 30S PIC[8] and increasing the affinity of IF2 to the 30S ribosomal subunit in the absence of initiator tRNA[9,10]. Increasing IF2 and initiator tRNA concentrations to 1 μM raised the fraction of simultaneous arrival to 45% while lowering the fraction of IF2 arriving first to 35% and the fraction of initiator tRNA arriving first to 10%. Thus, the order of IF2 and initiator tRNA arrival does not strictly follow a defined sequence, but is greatly affected by ligand concentrations and other initiation factors. Whereas at lower concentrations, the ligands arrive independently, simultaneous arrival of both ligands could be a more common mechanism at near physiological concentrations.

50S subunit joining to a 30S PIC to form a 70S initiation complex (70S IC) is the second major molecular event of initiation. To track subunit joining, we used 50S subunits labelled with single dyes, which were shown to be functional in prior intersubunit fluorescence resonance energy transfer (FRET) studies[11,12]. We delivered Cy5–50S subunits to immobilized 30S PICs (see Methods). IF2 in the presence of GTP drives rapid, stable subunit joining[8] (Fig. 2a). At 2 μM IF2–GTP, Cy5–50S subunits joined rapidly to 30S PICs with an observed on rate $k_{on} = 1 \times 10^6\ M^{-1}\ s^{-1}$ (corresponding to an exponential lifetime of $\tau = 9$ s), forming complexes whose lifetime was limited by photobleaching ($\tau = 38$ s) (Fig. 2b and Supplementary Fig. 4 and Supplementary Text 2). In accordance with previous studies[8], omitting IF2 resulted in slow and unstable subunit joining, decreasing k_{on} to $0.3 \times 10^6\ M^{-1}\ s^{-1}$ ($\tau = 29$ s) and 50S lifetime to $\tau = 6$ s. In the presence of IF2 and non-hydrolysable GDPNP, 50S subunit arrival rate was similar to that of IF2–GTP. However, 50S subunit stability decreased to a lifetime of $\tau = 28$ s, consistent with prior intersubunit FRET results that GDPNP-bound IF2 can guide stable subunit joining without GTP hydrolysis[11,13]. Addition of the other two initiation factors, IF1 and IF3, at 1 μM each to 2 μM IF2–GTP did not appreciably change the k_{on} or the lifetime of the 50S subunit on our model mRNA.

Subunit joining accelerates GTP hydrolysis by IF2, and IF2–GDP quickly dissociates from the ribosome[13]; elongator tRNA arrival finalizes transition into elongation. Yet the relative timings of IF2 release, 50S subunit joining and elongator tRNA binding are not known. To monitor these events in real time, we delivered Cy5–IF2, Cy3.5–50S and Phe-(Cy2)tRNA^{Phe} (as a ternary complex of tRNA, elongation factor EF-Tu and GTP, abbreviated as tRNA–EF-Tu–GTP; EF-Tu encoded by *tufA/B*) to 30S PIC loaded with fMet-(Cy3)tRNA^{fMet}, simultaneously tracking four different labelled components (see Methods). An IF2 signal was followed by rapid and stable subunit joining ($t > 10$ s) in the presence of GTP (Fig. 3a). IF2 with GTP bound yielded stable tRNA binding ($t > 1$ s) after 50S subunit joining; only brief tRNA sampling occurs with GDPNP.

Post-synchronizing the four-colour experiments with IF2–GTP to 50S arrival revealed an overlap between the IF2 and 50S signals of $\tau = 2$ s on a 70S IC (Fig. 3b and Supplementary Fig. 5 and Supplementary Text 3). This overlap time was independent of 50S subunit

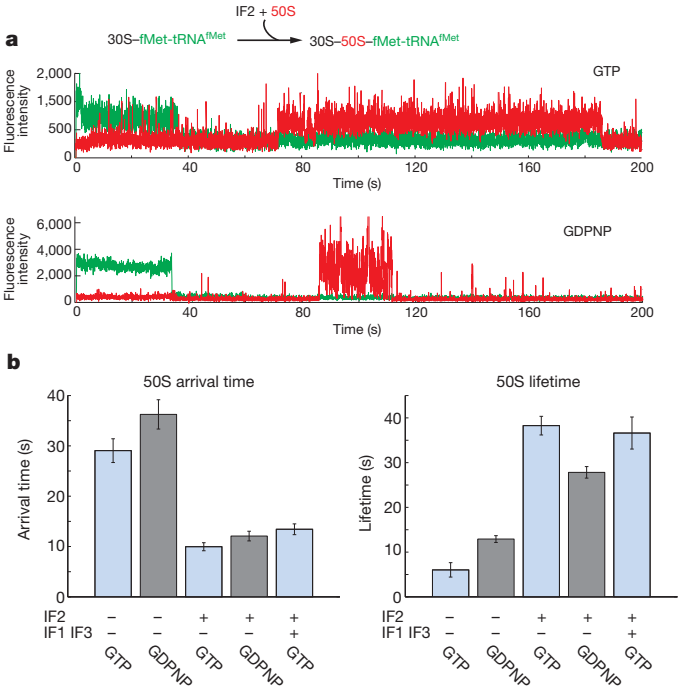


Figure 2 | 50S subunit joining to 30S PIC. a, The appearance of a stable Cy5 signal is used to identify the arrival of the 50S subunit (see Methods). The time until 50S arrival and the length of the 50S signal are then characterized. **b**, The arrival times of a 50S subunit to and the observed 50S subunit lifetimes on a 30S PIC are fitted to single-exponential functions and plotted with standard deviation (s.d.) error bars. The presence of IF2 is critical for efficiently and stably forming 70S complexes. From left to right for both panels, $n = 246$, $n = 283$, $n = 451$, $n = 262$ and $n = 253$.

concentration, suggesting that unimolecular processes occur within this overlap (Fig. 3c). During this period, IF2 rapidly hydrolyses GTP, rearranges the 70S IC, and then dissociates from the ribosome; consistent with this interpretation, the lifetime of IF2–GDP on 70S ribosomes was $\tau = 1.2$ s (Supplementary Fig. 2c). Interestingly, the arrival time of the elongator tRNA after subunit joining showed a similar lag of ~2 s. Increasing tRNA concentration beyond 200 nM had no statistically significant effect on tRNA arrival time, suggesting that tRNA arrival is not a rate-limiting step (Fig. 3c). The temporal correlation of IF2 dissociation and tRNA arrival during this 2 s window was not absolute; when single-molecule trajectories were post-synchronized to IF2 departure, tRNA arrival frequency increased after IF2 release, but ~20% of tRNA molecules arrived before IF2 departure.

Further analysis of these experiments explains how IF2 controls the transition into elongation. In the presence of IF2–GDPNP, very little elongator tRNA density was observed. The overall frequency of all elongator tRNA-binding events within a 2-s window before and after IF2 departure was similar in the presence of either GTP or GDPNP (Fig. 3d and Supplementary Text 4). However, the majority of tRNA arrival events were short-lived sampling events in GDPNP whereas most of the tRNA events in GTP involved stable (>1 s) binding, indicating that before GTP hydrolysis, only short-lived elongator tRNA sampling events are allowed.

Single-molecule techniques can distinguish among heterogeneous populations of molecules. By tracking individual dye-labelled molecules, we have shown that initiation does not follow a strictly linear mechanism whereby the translational machinery is rigidly assembled in a well-defined order. Although the ribosome must proceed through defined stages, such as forming the 30S PIC and 70S IC, there are numerous pathways available to it (Fig. 4).

During the formation of 30S PIC, IF2 and fMet-tRNA^{fMet} must bind to the 30S subunit to establish a reading frame on the mRNA and

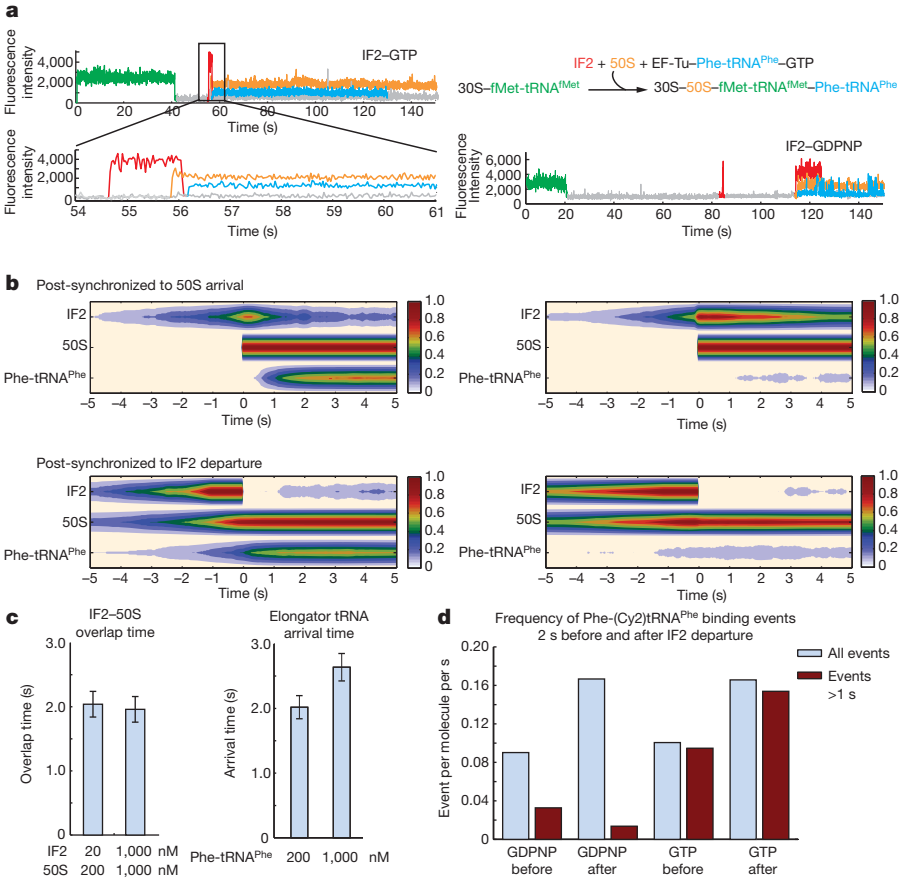


Figure 3 | Timing of IF2 departure and elongator tRNA arrival after 70S complex formation. **a**, See Methods for experimental setup. The timing of IF2 departure is determined by the disappearance of the Cy5 signal. **b**, The panels represent post-synchronization plots on 50S subunit arrival and IF2 departure at 1 μM dye-labelled ligand concentrations. With GTP ($n = 161$), there is a ~2 s overlap between the IF2 and 50S subunit signals with a strong elongator tRNA density. With GDPNP ($n = 87$), the overlap between IF2 and 50S is longer (~10 s) but there is little elongator tRNA density. **c**, The exponential lifetimes of the IF2-50S subunit overlap (left) in the presence of GTP did not depend on IF2 or 50S subunit concentrations. Increasing the elongator tRNA concentration also did not reduce the wait time until tRNA arrival (right). From left to right for both panels, $n = 169$ and $n = 161$; error bars are s.d. **d**, The event frequencies per molecule 2 s before and after IF2 departure are similar in both GTP ($n = 161$) and GDPNP ($n = 87$). Most events in GDPNP were removed by only counting events >1 s. Most of the events in GTP were longer-lived tRNA-binding events.

prime the 30S subunit for subunit joining. We observed all possible orders of binding occurring under different conditions. The binding pattern changes depending on concentrations of both molecules and on the presence of other initiation factors, and our results suggest that simultaneous arrival of IF2 and the tRNA may dominate *in vivo*.

The last stages of successful initiation ensure stable 70S ribosome assembly and configure it to accept the first elongator tRNA. IF2–GTP is required for stable 70S complex formation. GTP hydrolysis by IF2 occurs rapidly (30 ms) after subunit joining, but our data show a lag of 1–2 s before elongator tRNA arrival. During this period, IF2–GDP is bound to the 70S ribosome. Cryo-electron microscopy (cryo-EM) structures show that the IF2–GDP adopts a different conformation from the GTP form, and moves away from the GTPase activation centre[14]. Our data show that IF2 occupancy after GTP hydrolysis on the 70S complex hinders elongator tRNA arrival, consistent with cryo-EM maps of IF2–GDP on the 70S ribosome showing partial steric clash

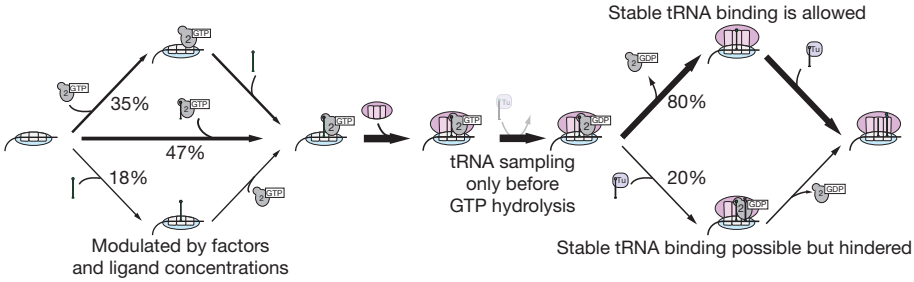


Figure 4 | The heterogeneous pathways of translation initiation. Multiple pathways are possible to reach the important stages of initiation as the ribosome converges to an elongation-competent 70S IC. Concentrations of initiation factors, tRNAs and ribosomal subunits all modulate the flux through the possible pathways that lead to successful initiation. After the 30S subunit binds to the mRNA, IF2 has a central role in channelling the ribosome towards elongation. At physiological concentrations of initiation factors and tRNAs, the majority of 30S PICs may be formed by IF2 bringing in the initiator tRNA to the 30S. IF2 also guides rapid and stable 50S subunit joining, while GTP hydrolysis by IF2 and its departure from the ribosome gates the stable binding of the first elongator tRNA.

with an incoming ternary complex (tRNA–EF-Tu–GTP) in the A site. These results suggest that IF2 release from the 70S complex controls the transition from initiation to elongation.

The single-molecule data presented here demonstrate the heterogeneous nature of translation initiation (Fig. 4). As the process evolves, the pathways converge to an elongation-competent 70S complex, with initiator tRNA positioned in the peptidyl-tRNA-binding site (P site), the correct intersubunit conformation, and IF2 clearance from the complex. Initiation factors guide the fidelity and timing of the process: IF1 and IF3 together regulate the order of IF2 and tRNA arrival and overall initiation efficiency, albeit via unclear mechanisms. IF2 guides the ribosome towards productive initiation and GTP hydrolysis governs transition into elongation. The single-molecule methods using ZMWs presented here can be broadly applicable to tracking compositional dynamics in other biological systems.

METHODS SUMMARY

Initiator and elongator tRNAs were dye labelled at the elbow position using Cy2-NHS (Cy2 conjugated to *N*-hydroxysuccinimide) or Cy3-maleimide dyes[15]; IF2 was labelled by cysteine (K791C) with Cy5-maleimide[16]. Ribosomal subunits were labelled using dye-conjugated oligonucleotide hybridization to mutant ribosomes[17] (Supplementary Fig. 6). Biochemical experiments confirmed the functionality of all labelled components[4,12]. Unless noted otherwise, all experiments were performed under buffer conditions described in Methods.

Data collection from ZMW chips was conducted using instrumentations and techniques described previously[3,4] (Supplementary Fig. 1). Fluorescence traces were recorded at 30 frames per second for 5 min, with delivery of ligands to start the experiment at $t = 7$ s. The photobleaching lifetimes of the labelled ligands were characterized as previously described[4] and were used to determine if a given loss of fluorescence signal is probably due to photobleaching or ligand dissociation. Data analysis on those traces was also conducted as described previously[4]. All error bars presented on figures show standard deviation errors from fitting the data to exponential decay functions.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 18 January; accepted 30 April 2012.
Published online 17 June 2012.

1. Laursen, B. S., Sorensen, H. P., Mortensen, K. K. & Sperling-Petersen, H. U. Initiation of protein synthesis in bacteria. *Microbiol. Mol. Biol. Rev.* **69,** 101–123 (2005).
2. Kozak, M. Initiation of translation in prokaryotes and eukaryotes. *Gene* **234,** 187–208 (1999).
3. Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299,** 682–686 (2003).
4. Uemura, S. *et al.* Real-time tRNA transit on single translating ribosomes at codon resolution. *Nature* **464,** 1012–1017 (2010).
5. Milon, P. *et al.* The ribosome-bound initiation factor 2 recruits initiator tRNA to the 30S initiation complex. *EMBO Rep.* **11,** 312–316 (2010).
6. Lockwood, A. H., Chakraborty, P. R. & Maitra, U. A complex between initiation factor IF2, guanosine triphosphate, and fMet-tRNA: an intermediate in initiation complex formation. *Proc. Natl Acad. Sci. USA* **68,** 3122–3126 (1971).
7. Petersen, H. U., Roll, T., Grunberg-Manago, M. & Clark, B. F. Specific interaction of initiation factor IF2 of *E. coli* with formylmethionyl-tRNA^{fMet}. *Biochem. Biophys. Res. Commun.* **91,** 1068–1074 (1979).
8. Antoun, A., Pavlov, M. Y., Lovmar, M. & Ehrenberg, M. How initiation factors tune the rate of initiation of protein synthesis in bacteria. *EMBO J.* **25,** 2539–2550 (2006).
9. Lockwood, A. H., Sarkar, P. & Maitra, U. Release of polypeptide chain initiation factor IF-2 during initiation complex formation. *Proc. Natl Acad. Sci. USA* **69,** 3602–3605 (1972).
10. Caserta, E. *et al.* Translation initiation factor IF2 interacts with the 30 S ribosomal subunit via two separate binding sites. *J. Mol. Biol.* **362,** 787–799 (2006).
11. Marshall, R. A., Aitken, C. E. & Puglisi, J. D. GTP hydrolysis by IF2 guides progression of the ribosome into elongation. *Mol. Cell* **35,** 37–47 (2009).
12. Aitken, C. E. & Puglisi, J. D. Following the intersubunit conformation of the ribosome during translation in real time. *Nature Struct. Mol. Biol.* **17,** 793–800 (2010).
13. Antoun, A., Pavlov, M. Y., Andersson, K., Tenson, T. & Ehrenberg, M. The roles of initiation factor 2 and guanosine triphosphate in initiation of protein synthesis. *EMBO J.* **22,** 5593–5601 (2003).
14. Myasnikov, A. G. *et al.* Conformational transition of initiation factor 2 from the GTP- to GDP-bound state visualized on the ribosome. *Nature Struct. Mol. Biol.* **12,** 1145–1149 (2005).
15. Blanchard, S. C., Kim, H. D., Gonzalez, R. L. Jr, Puglisi, J. D. & Chu, S. tRNA dynamics on the ribosome during translation. *Proc. Natl Acad. Sci. USA* **101,** 12893–12898 (2004).
16. Marshall, R. A. *Regulation of Protein Synthesis via Changes in Ribosome Conformation.* PhD thesis, Stanford Univ. (2008).
17. Dorywalska, M. *et al.* Site-specific labeling of the ribosome for single-molecule spectroscopy. *Nucleic Acids Res.* **33,** 182–189 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Supported by National Institutes of Health grant GM51266 (J.D.P.) and the Japan Science and Technology Agency (S.U.).

Author Contributions A.T., A.P. and S.U. conducted the experiments and performed the analysis; R.A.M. prepared and provided experimental materials; J.K. provided technical expertise with instrumentation and data processing; S.U. and J.D.P. designed experiments; and all authors discussed results and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details accompany the full-text HTML version of the paper at www.nature.com/nature. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to S.U. (s-uemura@gsc.riken.jp) or J.D.P. (puglisi@stanford.edu).

METHODS

Dye-labelled ligands. *E. coli* ribosomal subunits, initiation factors and elongation factors were prepared and purified as described[11,15,17,18]. $tRNA^{fMet}$ and $tRNA^{Phe}$ were labelled with fluorescent cyanine dyes at their elbow positions (U8 or U47), purified and aminoacylated as previously described[15,18]. A single-cysteine mutant of IF2 (C599A and K791C) was labelled with monomaleimide-Cy5 (GE Lifesciences) according to instructions from GE Lifesciences[16]. Supplementary Fig. 6 shows the location of dye on the biomolecules.

Experimental conditions. Ribosome initiation complexes were assembled at 0.25 μM 30S subunit concentration in a polymix buffer (50 mM Tris-acetate (pH 7.5), 100 mM potassium chloride, 5 mM ammonium acetate, 0.5 mM calcium acetate, 5 mM magnesium acetate, 0.5 mM EDTA, 5 mM putrescine-HCl and 1 mM spermidine) as described previously[4]. Nucleotide concentration is at 4 mM for GTP, GDP and GDPNP in all experiments.

30S PIC immobilization. Biotinylated mRNAs were used to immobilize 30S PICs with or without fMet-$tRNA^{fMet}$. Complexes were tethered to the biotin-PEG-derivatized quartz surface on the bottom of ZMW wells through a tetrameric neutravidin adaptor molecule by establishing PEG-biotin–neutravidin–biotinylated-mRNA complexes. The mRNA used contains the following in order from 5′ to 3′: a 5′ UTR and Shine–Dalgarno sequence derived from gene 32 of the T4 phage, an AUG start codon, 6 repeats of Phe-Lys codons, a UAA stop codon and 4 spacer Phe codons (Supplementary Fig. 1). Immobilized PICs were identified by initiator tRNA or 30S-subunit fluorescence and were distributed in ZMW holes according to Poisson statistics[4]. A single photobleaching step for the scoring dye confirms the single occupancy of the ZMW. Control experiments without mRNA demonstrated the absence of non-specific surface interactions at concentrations up to 1 μM of labelled tRNAs, factors or ribosomes. Thus, fluorescent events observed here represent true interactions of translation components with immobilized 30S subunits.

Observing the order of arrival of IF2 and initiator tRNA. We delivered a mixture of fMet-(Cy3)$tRNA^{fMet}$, Cy5–IF2–GTP and Cy3.5–50S to immobilized Alexa488–30S at 20 nM, 200 nM or 1,000 nM of each reagent. When present, IF1 and IF3 are at 1 μM. The appearance of a stable 50S signal was used to identify productive tRNA and IF2 binding events. The relative timing of IF2 and $tRNA^{fMet}$ arrival to the ribosome was determined by the order that their respective signals appear in each trace.

Observations for all experiments were done at 30 frames per second with ~33.3 ms exposure. Therefore, simultaneous arrival of fluorescence signals can be either genuine simultaneous arrival events or the two events happening sequentially within the exposure time. From the tRNA and IF2 arrival rates observed in control experiments, we calculated the percentage of apparent simultaneous events that can be attributed to sequential events happening in quick succession within one frame of exposure. We subtracted that part from the simultaneous events we observed and represented the subtracted portion as 'instrument uncertainty', or the grey portion on the bar graph in Fig. 1c.

Observing the role of IF2 in subunit joining. We delivered 200 nM of Cy5–50S to immobilized Cy3–30S at different magnesium concentrations (2.5–10 mM) both with and without 1 μM IF2 in either GTP or GDPNP. Where present, IF1 and IF3 are at 1 μM. The wait time until the appearance of the 50S signal and its lifetime was analysed to determine the efficiency of subunit joining under the different conditions.

Observing the relative timing of IF2 departure and elongator tRNA arrival. We delivered 20 to 1,000 nM Cy5–IF2, 200 to 1,000 nM Cy3.5–50S and 200 to 1,000 nM Phe–Cy2–$tRNA^{Phe}$ in a ternary complex with EF-Tu and GTP to 30S PIC loaded with fMet–Cy3–$tRNA^{fMet}$, simultaneously tracking four different labelled components in either GTP or GDPNP. We determined the overlap times of the IF2 signal with the 50S signal and the relative timing of IF2 departure to the arrival of the elongator tRNA, as well as the frequency of tRNA binding on the 70S ribosome after subunit joining.

18. Blanchard, S. C., Gonzalez, R. L., Kim, H. D., Chu, S. & Puglisi, J. D. tRNA selection and kinetic proofreading in translation. *Nature Struct. Mol. Biol.* **11,** 1008–1014 (2004).